DOCUMENT RESUME

ED 183 538 TM 010 107

AUTHOR	Stewart, Bruce L.
TITLE	Testing for Critical Thinking: A Review of the Resources. Rational Thinking Reports Number 2.
INSTITUTION	Illinois Univ., Urbana. Bureau of Educational Research.
PUB DATE	79
NOTE	148p.; For related document, see TM 010 106
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Cognitive Tests: *Critical Thinking; Test Construction: Test Reliability: *Test Reviews; Test Validity
IDENTIFIERS	Test Bibliographies

ABSTRACT

Tests for assessing critical thinking skills or abilities are reviewed in order to: (1) compare various measures and provide a critical review of each instrument: (2) encourage researchers to refine them: and (3) identify problems in measuring critical thinking by analyzing weaknesses in the tests. The publication is divided into two sections: a general discussion of the similarities and differences among the tests: and a review of each test discussed in the text, as well as a number of others not specifically mentioned. Information included as part of each test review consists of: test title, author, year of publication, and most recent edition: test description: sample test item: reliability and validity information: availability: comments: references to additional sources of information: and references to specific test citations. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183588

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

RECEIVED
ERIC/TM

TESTING FOR CRITICAL THINKING: A REVIEW OF THE RESOURCES

BY

BRUCE L. STEWART

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Bruce L. Stewart

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RATIONAL THINKING REPORTS
NUMBER 2

General Editor of Rational Thinking Reports:
Robert H. Ennis

Published by

The Illinois Rational Thinking Project

Bureau of Educational Research
University of Illinois at Urbana-Champaign
Urbana, Illinois
1979

TM010 107

Table of Contents

Page

Acknowledgements

Introduction . . . 1

General and aspect-specific critical thinking tests . . . 2

Discussion of instruments . . . 3

- 1. general tests . . . 4
- 2. aspect-specific tests . . . 8
- 3. testing formats . . . 11
- 4. target populations . . . 13
- 5. norm-referenced and criterion-referenced measures . . . 14
- 6. availability . . . 16
- 7. measurement problems . . . 17
    - a. single correct responses . . . 18
    - b. confounding of effects . . . 19
    - c. transferability . . . 20
    - d. test length; fidelity vs. bandwidth . . . 22

Summary . . . 24

References . . . 25

Appendix: Summaries and reviews of selected critical thinking tests . . . 26

- Assumption Recognition Index (I and II) . . . 27
- Class Reasoning Index . . . 30
- Conditional Reasoning Index . . . 33
- Cornell Class Reasoning Test, Form X . . . 36
- Cornell Conditional Reasoning Test, Form X . . . 39
- Cornell Critical Thinking Test, Level X . . . 42
- Cornell Critical Thinking Test, Level Z . . . 46

Table of Contents (cont.)

Page

Evaluation Aptitude Test 51

Inductive Reasoning Test 58

Interpretation of Data Test. 61

Illinois Conditional Reasoning Test. 68

Illinois Test of Ability to Judge Interpretations of Data. . . . 73

Logical Reasoning. 78

Logical Reasoning Test 83

Moorburg Letter Test 89

Ohio Thinking Checkup. 93

Pearson-Hyram Test of Critical Thinking. 98

Recognizing Reliable Observations. 103

Reflective Judgment Interview. 105

Science Process Skills Test. 109

Smith-Sturgeon Conditional Reasoning Test. 114

Strong and Weak Argument Test. 119

Test of Ability to Find Appropriate Illustrations. 122

Uncritical Inference Test. 125

Watson-Glaser Critical Thinking Appraisal. 129

ACKNOWLEDGEMENTS

This report could not have been completed without the support and assistance freely provided by members of the Illinois Rational Thinking Project and the Bureau of Educational Research at the University of Illinois, Urbana, Illinois. I am particularly indebted to Dr. Robert Ennis for his comments and criticisms on earlier drafts of this review. Steve Norris, Tom Tomko and Frances Wagner also offered thoughtful suggestions for improvements, and our many discussions of problems and issues related to critical thinking provided needed direction for this review. Finally, I wish to acknowledge the efforts of the secretarial staff of the Bureau of Educational Research; their thoroughness and hard work contributed significantly to the preparation of this manuscript.

Testing for Critical Thinking — A Review of the Resources

Bruce L. Stewart

Introduction

Many educators are currently concerned with the teaching and evaluation of critical thinking ability. Although the term does not have a single, standardized usage, it seems likely that most informed researchers and educators in the field would agree that critical thinking involves the reasonable evaluation of statements. Such statements may be of fact or opinion; oral or written; and may concern numerical as well as other types of data. A "reasonable evaluation" of these kinds of statements includes such processes as deciding what other statements (if any) follow logically from those given; assessing the nature and strength of justifications offered in support of particular statements; recognizing equivocation in the use of a word or words; and identifying unstated assumptions in a line of reasoning. Ennis (1962) proposes the general definition of critical thinking adopted in this paper: critical thinking is "...the correct assessing of statements" (p. 159). Interested readers are directed to this reference for a detailed explication of the concept, critical thinking, and the abilities which comprise it.

This review discusses a number of instruments purportedly measuring critical thinking ability or skills and abilities which might be reasonably subsumed under that title. The review was written for several purposes: (1) to compare and contrast measures designed to assess critical thinking ability, and to provide a critical review of each instrument; (2) to encourage the efforts of educators and educational researchers to refine existing measures or to develop additional instruments; and (3) to identify some of the difficulties involved in attempting to measure critical thinking ability by pinpointing both general and specific weaknesses in the tests. In general, the review attempts to acquaint readers with the types of resources typically used to measure critical thinking ability.

The first large section of the review is devoted to a general discussion of the similarities and differences of the instruments included in this review. The second portion contains reviews of each measure discussed in the text, as well as a number of others not specifically mentioned. No claim is made for the comprehensiveness of this review, since time limitations prevented the inclusion of additional instruments. However, in the judgment of this reviewer, the tests included here are fairly representative of the types of measures which have been developed to assess critical thinking ability. An attempt was also made to include tests which were widely-used measures which were innovative; in addition, both currently-available and out-of-print instruments are discussed. All comments are based upon a first-hand examination of each test by the reviewer.

General and aspect-specific critical thinking tests:

Critical thinking tests may be roughly divided into two types. A general critical thinking test is one which aims at an overall assessment of critical thinking ability, by sampling many (or possibly all) of the skills believed to make up the construct. In contrast, an aspect-specific critical thinking test is one designed to assess only one (or a few) components of critical thinking ability, such as the ability to judge observation statements or the ability to reason by the principles of conditional logic.

Both types of tests are included in this review. Obviously, however, whether a test is seen as one which is general or aspect-specific depends upon the definition of critical thinking ability adopted. A sufficiently narrow definition of critical thinking ability may restrict it to a single skill. If, for example, critical thinking ability were defined as the ability to reason in accord with the principles of conditional logic, then a test of this skill would qualify as a general critical thinking test. For the purposes of classifying instruments discussed in this review as either general or aspect-specific, the concept of critical thinking proposed by Ennis (1962) was adopted. Hence, some instruments which may claim to be general tests of critical thinking ability may not sample enough skills to be so considered under the definition of critical thinking ability used here.

Discussion of instruments:

One of the problems in comparing tests of critical thinking ability is that few definitions or explications of the construct are offered in test manuals. Among the instruments included here, only the Cornell Critical Thinking Test, Level X, the Cornell Critical Thinking Test, Level Z, and the Watson-Glaser Critical Thinking Appraisal are accompanied by manuals containing a substantial discussion of the concept of critical thinking ability which served as the basis for development of the measure. A number of test manuals neglected to include any statement of this nature at all; others devoted a general sentence or two to explaining what the test claimed to measure. Abbreviated or absent discussions of the concept of critical thinking ability underlying the test often entail that prospective test users must decide for themselves what traits or abilities are measured by the instrument. Probably the most common approach taken in the test manuals, however, was one of listing the traits or abilities purportedly measured by the test, with or without a subsequent short explanation. However, statements about what the test measured and the test titles themselves could be misleading. For example, the Evaluation Aptitude Test manual claims that the instrument measures "...a subject's composite aptitude to reason deductively" (p. 1), yet the test items are all class-logic syllogisms. Similarly, the Logical Reasoning Test (1955) is claimed by its authors to measure "...the ability to judge the logical soundness of meaningful conclusions," which they explicitly equate with critical thinking ability. The instrument, however, is composed only of class-reasoning items. In addition, test authors or publishers may adopt an unusual or idiosyncratic definition of critical thinking ability when designing a test. The danger here is that the meaning of scores on such measures may be misinterpreted by persons unfamiliar with the concept of critical thinking ability on which the instrument is based. Caution is therefore required in selecting any critical thinking test for possible use; this reviewer strongly recommends a first-hand examination of the test manual, the test itself, and any other relevant materials before using any test of critical thinking ability.

1. general tests:

Of the tests included here, very few appear to be general tests of critical thinking ability, although more instruments claim to be than actually seem to be general tests. No measure samples all the components of critical thinking identified by Ennis (1962). However, several instruments attempt to sample a fairly large number of abilities, such as the Cornell Critical

Thinking Test, Level X; the Cornell Critical Thinking Test, Level Z; the Logical Reasoning Test (1950); the Pearson-Hyram Test of Critical Thinking; and the Watson-Glaser Critical Thinking Appraisal. These instruments are usually composed of several subtests, each devoted to sampling one or a few of the abilities purportedly measured by the test. Although this format in some cases allows the use of part scores, (in the Logical Reasoning Test, for example, a total of 11 scores are computed) the manuals accompanying the various measures usually discourage the use of part scores. When the use of part scores is discouraged, the test manuals generally point out that such scores tend to be unreliable, and therefore should not be used as indicators of mastery of specific critical thinking skills. One reason for this recommendation against the use of part scores is that subtests frequently contain a limited number of items. Three subtests of the Cornell Critical Thinking Test, Level Z each contain only four items, for example. Although the items included in the subtests may actually sample the abilities intended, the reliability of each subtest can be expected to suffer somewhat, since test reliability is a function of test length. Shorter tests generally yield lower reliability estimates. Further, subtests with greater numbers of items are more likely to sample particular skills and abilities adequately, hence making scores on these subtests more meaningful. When part scores are sufficiently reliable to permit use, however, the utility of the instrument is often improved. Specific individual difficulties can be identified and used as a basis for remedial instruction. The use of part scores should also be a matter of individual judgment; in the opinion of this reviewer, prospective test users should consider not only the reliability of such part scores, but also whether particular subtests are likely to provide the kind of information desired.

General tests also differ greatly in the abilities they claim to sample. Some instruments (such as the Logical Reasoning Test (1950) and the Watson-Glaser Critical Thinking Appraisal, among others) do not sample semantic skills such as identifying equivocation or ambiguity in a line of reasoning. Tests also differ in the emphasis given to assessing particular abilities; the Watson-Glaser Critical Thinking Appraisal includes 25 items intended to test for the ability to reason deductively (class and conditional logic), whereas the Cornell Critical Thinking Test, Level Z uses only 10 items for the same purpose. Despite its title, the Logical Reasoning Test (1950) does not sample conditional or class reasoning ability at all. These differences in the abilities sampled and the relative intensity with which they are sampled may be partly accounted for by differing concepts of critical thinking ability underlying the construction of each test.

However, among the general critical thinking tests cited here, the three currently-available instruments do overlap in the abilities they are intended to assess. The Cornell Critical Thinking Test, Level X, the Cornell Critical Thinking Test, Level Z, and the Watson-Glaser Critical Thinking Appraisal each attempt some assessment of deductive reasoning ability; ability to evaluate evidence offered in support of a statement; and ability to recognize unstated assumptions in a line of reasoning. Further, the Cornell Critical Thinking Test, Level Z and the Watson-Glaser Critical Thinking Appraisal both include subtests designed to assess an examinee's ability to evaluate arguments. The two Cornell tests also share sections on judging the reliability of statements. Despite what may appear to be considerable overlap, however, these instruments should not be assumed to be equivalent. Each of these measures samples particular abilities somewhat differently, and with different degrees of emphasis.

The instruments mentioned earlier were judged by this reviewer to represent attempts to provide general assessments of critical thinking ability by sampling a broad range of skills. A number of other measures also claimed or appeared constructed to yield a general appraisal of critical thinking ability, although that label was not always applied to the skill(s) purportedly measured by the test. These tests included the Evaluation Aptitude Test; the Reflective Judgment Interview; the Logical Reasoning Test (1955); and the Moorburg Letter Test. For the most part, however, these measures exhibit one or both of the following characteristics, which may suggest that they fail to provide an accurate general assessment of individual critical thinking ability: (1) only a few (usually no more than three) skills or abilities were specifically identified, either in the manual or on the test itself, as being sampled by the test. Hence, it seems unlikely that measures sampling only a few abilities would be adequate general indicators of critical thinking ability if the concept of critical thinking proposed by Ennis (1962) is adopted as a working definition. (2) Specific skills or abilities are often sampled by only a few items. Such limited sampling contributes to the unreliability of part scores, and may not provide an adequate sample of the intended skill. This topic is discussed in greater detail in the section of this review devoted to problems of measurement. Although the instruments mentioned above each have specific strengths and weaknesses, they are all less broadly-based than the general critical thinking tests cited earlier.

2. aspect-specific tests:

The remaining tests included in this review are all aspect-specific; that is, they attempt to provide an assessment of proficiency in a particular skill (which may or may not be composed of other sub-skills). Each skill or ability evaluated by an aspect-specific test might reasonably be included as a component of critical thinking ability. Even the concept of critical thinking ability proposed by Ennis (1962) however, does not specifically include all the abilities represented by the group of aspect-specific tests taken as a whole.

Aspect-specific tests potentially provide a more thorough sampling (and hence, a more accurate assessment) of an individual's performance with respect to a particular skill or ability. By focusing on a single ability, aspect-specific tests can sample that ability or skill more intensively, and under a broader range of conditions. A measure of conditional reasoning ability such as the Cornell Conditional Reasoning Test, for example, attempts to assess this ability by including test items which require mastery of different conditional reasoning principles for their solution; items which vary in their content; and items which require examinees to apply the basic principles of conditional reasoning in combination in order to arrive at a solution. Other measures, such as the Conditional Reasoning Index, employ items using both emotional and neutral language in order to assess the examinee's conditional reasoning ability.

And many aspect-specific tests also provide part scores. The Conditional Reasoning Index generates scores on emotionally-laden as well as neutral items, in addition to a total test score. All of the instruments claiming to measure class or conditional reasoning ability yield part scores for

Testing for Critical Thinking — A Review of the Resources

Bruce L. Stewart

Introduction

Many educators are currently concerned with the teaching and evaluation of critical thinking ability. Although the term does not have a single, standardized usage, it seems likely that most informed researchers and educators in the field would agree that critical thinking involves the reasonable evaluation of statements. Such statements may be of fact or opinion; oral or written; and may concern numerical as well as other types of data. A "reasonable evaluation" of these kinds of statements includes such processes as deciding what other statements (if any) follow logically from those given; assessing the nature and strength of justifications offered in support of particular statements; recognizing equivocation in the use of a word or words; and identifying unstated assumptions in a line of reasoning. Ennis (1962) proposes the general definition of critical thinking adopted in this paper: critical thinking is "...the correct assessing of statements" (p. 159). Interested readers are directed to this reference for a detailed explication of the concept, critical thinking, and the abilities which comprise it.

This review discusses a number of instruments purportedly measuring critical thinking ability or skills and abilities which might be reasonably subsumed under that title. The review was written for several purposes: (1) to compare and contrast measures designed to assess critical thinking ability, and to provide a critical review of each instrument; (2) to encourage the efforts of educators and educational researchers to refine existing measures or to develop additional instruments; and (3) to identify some of the difficulties involved in attempting to measure critical thinking ability by pinpointing both general and specific weaknesses in the tests. In general, the review attempts to acquaint readers with the types of resources typically used to measure critical thinking ability.

The first large section of the review is devoted to a general discussion of the similarities and differences of the instruments included in this review. The second portion contains reviews of each measure discussed in the text, as well as a number of others not specifically mentioned. No claim is made for the comprehensiveness of this review, since time limitations prevented the inclusion of additional instruments. However, in the judgment of this reviewer, the tests included here are fairly representative of the types of measures which have been developed to assess critical thinking ability. An attempt was also made to include tests which were widely-used measures which were innovative; in addition, both currently-available and out-of-print instruments are discussed. All comments are based upon a first-hand examination of each test by the reviewer.

General and aspect-specific critical thinking tests:

Critical thinking tests may be roughly divided into two types. A general critical thinking test is one which aims at an overall assessment of critical thinking ability, by sampling many (or possibly all) of the skills believed to make up the construct. In contrast, an aspect-specific critical thinking test is one designed to assess only one (or a few) components of critical thinking ability, such as the ability to judge observation statements or the ability to reason by the principles of conditional logic.

Both types of tests are included in this review. Obviously, however, whether a test is seen as one which is general or aspect-specific depends upon the definition of critical thinking ability adopted. A sufficiently narrow definition of critical thinking ability may restrict it to a single skill. If, for example, critical thinking ability were defined as the ability to reason in accord with the principles of conditional logic, then a test of this skill would qualify as a general critical thinking test. For the purposes of classifying instruments discussed in this review as either general or aspect-specific, the concept of critical thinking proposed by Ennis (1962) was adopted. Hence, some instruments which may claim to be general tests of critical thinking ability may not sample enough skills to be so considered under the definition of critical thinking ability used here.

Discussion of instruments:

One of the problems in comparing tests of critical thinking ability is that few definitions or explications of the construct are offered in test manuals. Among the instruments included here, only the Cornell Critical Thinking Test, Level X, the Cornell Critical Thinking Test, Level Z, and the Watson-Glaser Critical Thinking Appraisal are accompanied by manuals containing a substantial discussion of the concept of critical thinking ability which served as the basis for development of the measure. A number of test manuals neglected to include any statement of this nature at all; others devoted a general sentence or two to explaining what the test claimed to measure. Abbreviated or absent discussions of the concept of critical thinking ability underlying the test often entail that prospective test users must decide for themselves what traits or abilities are measured by the instrument. Probably the most common approach taken in the test manuals, however, was one of listing the traits or abilities purportedly measured by the test, with or without a subsequent short explanation. However, statements about what the test measured and the test titles themselves could be misleading. For example, the Evaluation Aptitude Test manual claims that the instrument measures "...a subject's composite aptitude to reason deductively" (p. 1), yet the test items are all class-logic syllogisms. Similarly, the Logical Reasoning Test (1955) is claimed by its authors to measure "...the ability to judge the logical soundness of meaningful conclusions," which they explicitly equate with critical thinking ability. The instrument, however, is composed only of class-reasoning items. In addition, test authors or publishers may adopt an unusual or idiosyncratic definition of critical thinking ability when designing a test. The danger here is that the meaning of scores on such measures may be misinterpreted by persons unfamiliar with the concept of critical thinking ability on which the instrument is based. Caution is therefore required in selecting any critical thinking test for possible use; this reviewer strongly recommends a first-hand examination of the test manual, the test itself, and any other relevant materials before using any test of critical thinking ability.

1. general tests:

Of the tests included here, very few appear to be general tests of critical thinking ability, although more instruments claim to be than actually seem to be general tests. No measure samples all the components of critical thinking identified by Ennis (1962). However, several instruments attempt to sample a fairly large number of abilities, such as the Cornell Critical

Thinking Test, Level X; the Cornell Critical Thinking Test, Level Z; the Logical Reasoning Test (1950); the Pearson-Hyram Test of Critical Thinking; and the Watson-Glaser Critical Thinking Appraisal. These instruments are usually composed of several subtests, each devoted to sampling one or a few of the abilities purportedly measured by the test. Although this format in some cases allows the use of part scores, (in the Logical Reasoning Test, for example, a total of 11 scores are computed) the manuals accompanying the various measures usually discourage the use of part scores. When the use of part scores is discouraged, the test manuals generally point out that such scores tend to be unreliable, and therefore should not be used as indicators of mastery of specific critical thinking skills. One reason for this recommendation against the use of part scores is that subtests frequently contain a limited number of items. Three subtests of the Cornell Critical Thinking Test, Level Z each contain only four items, for example. Although the items included in the subtests may actually sample the abilities intended, the reliability of each subtest can be expected to suffer somewhat, since test reliability is a function of test length. Shorter tests generally yield lower reliability estimates. Further, subtests with greater numbers of items are more likely to sample particular skills and abilities adequately, hence making scores on these subtests more meaningful. When part scores are sufficiently reliable to permit use, however, the utility of the instrument is often improved. Specific individual difficulties can be identified and used as a basis for remedial instruction. The use of part scores should also be a matter of individual judgment; in the opinion of this reviewer, prospective test users should consider not only the reliability of such part scores, but also whether particular subtests are likely to provide the kind of information desired.

General tests also differ greatly in the abilities they claim to sample. Some instruments (such as the Logical Reasoning Test (1950) and the Watson-Glaser Critical Thinking Appraisal, among others) do not sample semantic skills such as identifying equivocation or ambiguity in a line of reasoning. Tests also differ in the emphasis given to assessing particular abilities; the Watson-Glaser Critical Thinking Appraisal includes 25 items intended to test for the ability to reason deductively (class and conditional logic), whereas the Cornell Critical Thinking Test, Level Z uses only 10 items for the same purpose. Despite its title, the Logical Reasoning Test (1950) does not sample conditional or class reasoning ability at all. These differences in the abilities sampled and the relative intensity with which they are sampled may be partly accounted for by differing concepts of critical thinking ability underlying the construction of each test.

However, among the general critical thinking tests cited here, the three currently-available instruments do overlap in the abilities they are intended to assess. The Cornell Critical Thinking Test, Level X, the Cornell Critical Thinking Test, Level Z, and the Watson-Glaser Critical Thinking Appraisal each attempt some assessment of deductive reasoning ability; ability to evaluate evidence offered in support of a statement; and ability to recognize unstated assumptions in a line of reasoning. Further, the Cornell Critical Thinking Test, Level Z and the Watson-Glaser Critical Thinking Appraisal both include subtests designed to assess an examinee's ability to evaluate arguments. The two Cornell tests also share sections on judging the reliability of statements. Despite what may appear to be considerable overlap, however, these instruments should not be assumed to be equivalent. Each of these measures samples particular abilities somewhat differently, and with different degrees of emphasis.

The instruments mentioned earlier were judged by this reviewer to represent attempts to provide general assessments of critical thinking ability by sampling a broad range of skills. A number of other measures also claimed or appeared constructed to yield a general appraisal of critical thinking ability, although that label was not always applied to the skill(s) purportedly measured by the test. These tests included the Evaluation Aptitude Test; the Reflective Judgment Interview; the Logical Reasoning Test (1955); and the Moorburg Letter Test. For the most part, however, these measures exhibit one or both of the following characteristics, which may suggest that they fail to provide an accurate general assessment of individual critical thinking ability: (1) only a few (usually no more than three) skills or abilities were specifically identified, either in the manual or on the test itself, as being sampled by the test. Hence, it seems unlikely that measures sampling only a few abilities would be adequate general indicators of critical thinking ability if the concept of critical thinking proposed by Ennis (1962) is adopted as a working definition. (2) Specific skills or abilities are often sampled by only a few items. Such limited sampling contributes to the unreliability of part scores, and may not provide an adequate sample of the intended skill. This topic is discussed in greater detail in the section of this review devoted to problems of measurement. Although the instruments mentioned above each have specific strengths and weaknesses, they are all less broadly-based than the general critical thinking tests cited earlier.

2. aspect-specific tests:

The remaining tests included in this review are all aspect-specific; that is, they attempt to provide an assessment of proficiency in a particular skill (which may or may not be composed of other sub-skills). Each skill or ability evaluated by an aspect-specific test might reasonably be included as a component of critical thinking ability. Even the concept of critical thinking ability proposed by Ennis (1962) however, does not specifically include all the abilities represented by the group of aspect-specific tests taken as a whole.

Aspect-specific tests potentially provide a more thorough sampling (and hence, a more accurate assessment) of an individual's performance with respect to a particular skill or ability. By focusing on a single ability, aspect-specific tests can sample that ability or skill more intensively, and under a broader range of conditions. A measure of conditional reasoning ability such as the Cornell Conditional Reasoning Test, for example, attempts to assess this ability by including test items which require mastery of different conditional reasoning principles for their solution; items which vary in their content; and items which require examinees to apply the basic principles of conditional reasoning in combination in order to arrive at a solution. Other measures, such as the Conditional Reasoning Index, employ items using both emotional and neutral language in order to assess the examinee's conditional reasoning ability.

And many aspect-specific tests also provide part scores. The Conditional Reasoning Index generates scores on emotionally-laden as well as neutral items, in addition to a total test score. All of the instruments claiming to measure class or conditional reasoning ability yield part scores for mastery of the particular reasoning principles included in the test. The Science Process Skills Test is composed of three subtests each intended to sample a different ability; examinees receive three part scores as well as a total test score. One advantage to the use of part scores on aspect-specific measures is the precision they potentially provide, which should prove useful for individual diagnostic and remedial purposes. Of course, the use of part scores also carries the same hazards on aspect-specific tests as on general tests of critical thinking ability. Both their reliability and the nature of the information part scores are designed to provide need to be evaluated by prospective test users.

What aspect-specific measures potentially gain in precision, they lose in generality, however. The aspect-specific tests included here vary in the comprehensiveness with which they sample single abilities, but none of these measures were intended to provide a general assessment of an individual's critical thinking ability. In addition, measures which could be expected to assess the same skill(s) often differ considerably from one another. The Conditional Reasoning Index, the Illinois Conditional Reasoning Test, the Cornell Conditional Reasoning Test and the Smith-Sturgeon Conditional Reasoning Test all claim to assess conditional reasoning ability. Their formats, number of conditional reasoning principles sampled, scoring methods and type of information they are designed to provide vary greatly. While a certain diversity among these instruments might be expected, undoubtedly some tests will be better suited to the particular requirements of prospective test users than others. Although these measures claim to assess the same ability, they should not be assumed to yield equivalent appraisals of individual abilities, or to sample conditional reasoning ability to the same extent or by comparable methods.

The aspect-specific tests included in this review cover a broad spectrum of skills and abilities. The traits measured range from deductive skills (class and conditional reasoning) to non-deductive skills (inductive reasoning; argument evaluation). One instrument, the Reflective Judgment Interview, seems designed to reveal an individual's epistemological and ethical perspective, and might be considered an affective measure of some traits or attitudes related to critical thinking ability (willingness to suspend judgment, for example). At least one other candidate for a component of critical thinking ability is excluded from those sampled by the measures cited in this review. Analogical reasoning ability — roughly, the ability to perceive ordered relationships among objects, words or numbers -- is not assessed directly by any of the tests included in this review, although tests (or subtests) of inductive reasoning ability may also sample analogical reasoning to some extent. Analogical reasoning problems are often included in standardized tests such as the Graduate Record Exam and the Miller Analogies Test; whether analogical reasoning should be considered a component of critical thinking, however, is probably a matter for both empirical and philosophical research into the nature of critical thinking ability.

Some combinations of the aspect-specific tests included in this review might serve as the basis for development of general critical thinking tests. Taken together, the Assumption Recognition Index (I and II), the Class Reasoning Index, the Conditional Reasoning Index and Recognizing Reliable Observations (all developed by the Instructional Objectives Exchange) might constitute a general, criterion-referenced critical thinking test. This use of aspect-specific tests should not be overlooked. Aspect-specific tests can serve as the basis for subtest development on a general measure;

the various aspect-specific tests included here should provide considerable flexibility in tailoring a general critical thinking test to particular needs.

3. testing formats:

Both general and aspect-specific tests of critical thinking employ several different testing formats. The standard, multiple-choice, paper-and-pencil written exam is the most common format of the instruments included here. Generally, tests require examinees to select an appropriate response (usually only one) from among those provided; most test items include at least two distractors. However, several measures depart from this pattern to one extent or another. The Assumption Recognition Index (I and II) requires examinees to select one of two alternative responses. Both the Strong and Weak Argument Test and the evaluation of arguments subtest on the Watson-Glaser Critical Thinking Appraisal also follow this format. One possible difficulty with this system is that the level of correct performance expected by chance increases to 50%, which could influence the interpretations to be placed on scores for these measures.

Other instruments depart more radically from the multiple-choice, paper-and-pencil testing format. The Science Process Skills Test requires that the examiner read the test directions aloud to examinees and that examinees wait for directions after completing each problem, insuring that everyone remains together. The test items are also constructed using illustrations rather than written statements, in order to minimize the effect of verbal ability on test performance. The Moorburg Letter Test does not consist of distinct items; rather, examinees are required to evaluate a fictitious letter to the editor and receive credit for identifying both fallacious and legitimate reasoning included in the letter. Their responses are given in essay format, and are evaluated against a special scoring key which also awards credit for the justifications offered by examinees for their responses. The Illinois Conditional Reasoning Test, the Smith-Sturgeon Conditional Reasoning Test and the Reflective Judgment Interview employ an individual interview format, in which the examinee provides his or her responses verbally. The two conditional reasoning measures also use actual manipulations of objects in test administration, and all these measures require scoring by special systems rather than the common "number correct" formula.

Both the multiple-choice, paper-and-pencil format and the less conventional alternatives have advantages. Multiple-choice tests are generally easily-administered and scored. In addition, this format is familiar to most examinees and test administrators, which may help to minimize errors due to misunderstood directions or novel required responses (for example, asking examinees to mark all alternatives they believe to be incorrect). In general, the less conventional formats are intended to provide a more accurate assessment of individual abilities, often by considering both the examinee's response and the justification offered for the response. These instruments attempt to distinguish between examinees who respond correctly by chance, and those who have mastered the skills considered necessary to produce a correct response. In contrast, multiple-choice tests usually are scored on the assumption that a correct response signifies knowledge. If a correction-for-guessing formula is employed, it still fails to discriminate those items answered correctly by chance from those answered correctly because the examinee had mastered the required skills. The penalty, however, for the potentially more accurate assessment of individual abilities provided by less conventional measures is usually one of time. Unconventional tests usually require increased administration and scoring time. Some require individualized administration as well. Most have scoring systems more complicated even than conventional correction-for-guessing formula scoring. These drawbacks may well make less conventional tests impractical for many prospective users.

4. target populations:

The majority of both general and aspect-specific critical thinking tests included here are intended for use with students in grade 7 and above. A few measures, however, specifically include lower grades in the groups for which the test is intended to be appropriate. Such measures include the Cornell Class Reasoning Test; the Cornell Conditional Reasoning Test; the Illinois Conditional Reasoning Test; the Smith-Sturgeon Conditional reasoning Test; and the Science Process Skills Test. The Ohio Thinking Checkup is particularly noteworthy in this context because of all the instruments included in this review, it probably comes closest to being a general test of critical thinking ability specifically designed for use in lower grades (4-6).

Several reasons may account for the apparent scarcity of critical thinking tests (aspect-specific or general) designed for use at lower grade levels. Critical thinking ability may include cognitive processes beyond the capabilities of most lower grade students; measurement efforts might be directed toward groups theoretically able to demonstrate behaviors taken as evidence of critical thinking ability. In addition, because many tests of critical thinking rely on the verbal skills of the examinee (vocabulary, reading ability), these tests may be inappropriate for use with groups in which such skills are minimally developed. The Science Process Skills Test represents one approach to minimizing the effects of verbal ability on critical thinking test scores by using illustrations and verbal directions to supplement written instructions whenever possible. The intent here is to compensate for pre-existing differences in verbal ability, and to avoid confounding the effects of verbal ability differences with differences due to the abilities the test is designed to measure. Another approach utilized by several of the measures included here is the individual oral interview, which allows examinee responses to be probed as an aid to interpretation and scoring. These sorts of difficulties are treated in more detail in a later discussion of general measurement problems faced by tests of critical thinking.

5. norm-referenced vs. criterion-referenced measures:

It should be noted at the outset of the following discussion that the classification of an instrument as norm-referenced or criterion-referenced is not intended to reflect a strict dichotomy. Without entering into all of the measurement issues involved here, it seems important to recognize that these types of tests usually differ in the principles by which they were constructed, the purposes for which they are best suited, and the kinds of interpretations which are applied to scores. Whether one type or the other is more appropriate for the particular purposes of a prospective test user is a matter for personal judgment. Readers unfamiliar with the field of criterion-referenced testing are referred to Popham (1978) for an introductory discussion of norm- and criterion-referenced tests and their differences.

Several tests included in this review specifically claim to be criterion-referenced measures. These are the Assumption Recognition Index (I and II); the Class Reasoning Index; the Conditional Reasoning Index; and Recognizing Reliable Observations. All of these tests were developed by the Instructional Objectives Exchange (IOX), and are fairly recent (1971). All are aspect-specific. These criterion-referenced tests differ from the other tests cited in this review in that scores on the criterion-referenced measures are not intended to be interpreted with reference to the performance of a larger group to which the test was also administered (the norm group). No percentiles, stanines, or other traditional normative data is included in the test manuals.

In contrast, many of the manuals of the remaining tests in this review provide normative information by which scores on these instruments are intended to be evaluated. However, two points are important in this context: (1) when normative data is either missing from the manual or no manual exists, and in the absence of specific statements in the test or the manual, the classification of an instrument as norm-referenced or criterion-referenced cannot be made reliably by inspection; (2) much of the normative data included in test manuals may be out-of-date, based on small samples, or otherwise inappropriate. Prospective test users are cautioned to examine all normative data carefully before using this information as a basis for score interpretation.

Although the nature and extent of normative data provided in the manuals of the tests included in this review varies widely, a few general trends are apparent. No test manual provided a specific date on which the normative group(s) was tested; most manuals give only a general indication, such as a range of years, of when normative information was compiled. Further, no test manual examined in this review provided fall, winter and spring norms. When provided, norms were given for a single point in time, often vaguely specified, as noted above. Precise specification of normative testing dates has important implications when gain scores on norm-referenced tests are to be used as the basis for educational evaluation. Horst, Tallmadge and Wood (1975) discuss this point. And finally, characteristics of the normative sample(s) were frequently not reported in any detail. A typical description of one normative sample might be "college seniors at an Ivy-League school." A more complete description of the normative sample(s) might aid prospective test users in evaluating the appropriateness for their purposes of the norms provided.

6. availability:

Selection of any critical thinking measure for possible use obviously depends upon the availability of the instrument -- and here the choices of prospective test users become somewhat limited. Only three currently-available general critical thinking tests were located for inclusion in this review. The Cornell Critical Thinking Test, Level X, the Cornell Critical Thinking Test, Level Z, and the Watson-Glaser Critical Thinking Appraisal are presently the most widely-used critical thinking measures. In contrast, many more aspect-specific tests are available than general tests. However, some of the measures cited in this review were developed as part of dissertation research or a published study, and have not been widely distributed. Although these instruments are in many cases available for inspection, they have no history of validated use. Often, such measures were included in this review because they represent innovative attempts to assess critical thinking ability or some aspect.

The greater availability of aspect-specific tests is undoubtedly a product of several factors. In many cases, the aspects measured by aspect-specific tests are of interest in their own right, such as class or conditional reasoning ability, for example. Neither is the investigation of critical thinking ability widely recognized as a distinct field of inquiry. This situation may be due in part to general unfamiliarity with the term "critical thinking ability", and the lack of consensus in its use. The three general tests noted above, for example, are indexed in the most recent Mental Measurements Yearbook under "study skills" and "reading". No separate classification for critical thinking measures is provided.

Additional critical thinking measures may become available, however. Recent concern for testing reasoning processes, as well as traditional verbal and quantitative skills resulted in the addition of a third section to the Graduate Record Examination. The latest section is designed to evaluate the analytical reasoning abilities of examinees, and is similar in many respects to some general critical thinking tests. Continued recognition of critical thinking ability as an important educational concern may encourage additional development and revision of both aspect-specific as well as general critical thinking tests.

7. measurement problems:

Increased test development and revision, however, will face many of the difficulties which confront the tests included here. The following discussions are intended to identify some of the most common measurement issues which must be dealt with in the construction, administration and interpretation of critical thinking instruments. Although many more could be cited, the measurement issues and problems discussed below are particularly important in attempting to assess critical thinking ability.

a. single correct responses:

Many test items often permit more than a single correct response, although only one response is keyed as correct. Tests designed to assess any sort of inductive reasoning abilities are particularly susceptible to this difficulty. Both the Inductive Reasoning Test and the first subtest (induction) of the Cornell Critical Thinking Test, Level X, suffer from this problem (although for different reasons). But a considerable number of other tests contain items which admit of more than one correct answer. This situation often indicates that the keyed answer was derived under particular assumptions; alternative, equally-plausible assumptions may justify other responses. In some cases, test directions or item construction is faulty.

The problem of alternative assumptions justifying additional responses stems from the nature of critical thinking itself. Because critical thinking is essentially a process of evaluation, different individuals may assess identical information differently, and arrive at different, but equally-justifiable conclusions. Test items which are intended to allow only one correct response fail to take account of this feature of the ability being measured. Further, such items usually reveal little of the processes used by an examinee in selecting a response. A correct response may not always indicate that an examinee thought critically in selecting it. The related problems of equally justifiable responses and determining if critical thinking has occurred in response selection have led some test designers to require examinees to provide a justification in addition to their response, which are then evaluated jointly. Hopefully, a more accurate assessment of an individual's ability results; at the very least, requiring both a response and a justification may help to minimize credit being given for correct responses due to chance. The fact that critical thinking tests attempt to assess a process, rather than only the final product (answer), complicates the problem of accurate measurement. At present, this difficulty appears significant for any test in which a correct response is not taken as conclusive evidence of particular mental processes.

b. confounding of effects:

Accurate interpretation of scores on any mental ability test depends upon being able to identify, to the extent possible, the factors responsible for test performance and their relative contributions to that performance. Ideally, individual differences in test performance are fully accounted for by differences in the ability being measured, plus measurement error. In practice, however, this ideal is not met. Individual test performance differences on measures of critical thinking ability may be due to actual ability differences, chance, differences in verbal ability (reading, vocabulary), fatigue, or many other factors.

One approach to improving the accuracy of score interpretation is represented by the Science Process Skills Test, which attempts to minimize the influence of factors other than actual differences in the abilities of concern on test performance. Differences in verbal ability are compensated for by the use of oral directions and illustrations whenever possible. This approach recognizes pre-existing individual differences in ability which may influence test performance, and attempts to insure that all examinees function at or above the minimum level required to understand and complete the test.

An alternative approach to improving the accuracy of test score interpretation is represented by the Illinois Test of Ability to Judge Interpretations of Data. The measure includes a "data reading" score, intended to identify individuals who have difficulty in comprehending information presented graphically or numerically. The test authors consider the ability to understand graphic and numerical data to be a prerequisite skill for data evaluation; hence, a low data reading score will tend to invalidate the other scores on the instrument for sine examinees. Data reading scores are intended to eliminate at least one alternative explanation for test performance. This approach attempts to identify, but not to compensate for, the influence of a relevant factor.

Other approaches to improving the interpretability of critical thinking test scores are no doubt possible. Yet there will always be some element of inference in going from test performance to assessment of an individual mental trait or ability. Improving the interpretability of test scores is likely to remain an important measurement and test development concern for tests of critical thinking as well as other mental abilities.

c. transferability:

Critical thinking skills are often claimed to be useful in everyday contexts. Few of the measures included in this review, however, provide testing contexts at all similar to the real-world contexts in which critical thinking skills might be utilized. The Moorburg Letter Test probably provides the most realistic testing context of any test in this review; the instrument is constructed as a letter-to-the-editor and contains sound as well as fallacious reasoning offered in support of the writer's position. Hopefully, the close parallel between the testing context and the actual circumstances in which critical thinking skills might be exercised provides one reason for believing that the test results will transfer to contexts outside of the testing situation. The Reflective Judgment Interview, the Illinois Conditional Reasoning Test and the Smith-Sturgeon Conditional Reasoning Test are also worth noting here. Each of these instruments is administered orally, and examinees are required to provide justification for their responses. This format parallels the flow of dialogue in a discussion, during which evidence and reasons are presented, challenged, examined and evaluated. Performance on these measures might, therefore, be taken as an indication of performance in actual, real-world contexts.

But this potentially greater transferability of results is bought at the cost of increased administration and scoring time. All of the measures providing more realistic testing contexts require special scoring techniques, and are not amenable to machine scoring. The Moorburg Letter Test is given in essay format, with all of the difficulties attending that format. In addition, essay tests are not easily compared to standard multiple-choice measures when psychometric properties of the instruments are used as a basis for comparison. Individual, oral test administration is both time-consuming and costly; these two drawbacks may render such measures impractical for a majority of prospective users.

Yet it seems likely that the question of transferability of test results will remain important as long as claims about the usefulness of critical thinking skills are not restricted to the testing context. And though predictive validity studies might provide some evidence that skills demonstrated on the test transfer to other contexts, such evidence is weak. For the criterion cannot be a similar test if any information concerning the transferability of results to non-testing contexts is to be gained. Ideally, criterion behaviors should be those that the test is designed to measure; unfortunately, such behaviors are often difficult to observe or quantify. Hence, one approach to insuring that test results will transfer to non-testing contexts is that of designing the testing context to parallel the actual context as closely as possible. The problem of transferability is common to many tests of mental abilities, yet it seems particularly important for critical thinking testing. Critical thinking skills are not subject-matter specific. Their wide applicability suggests that tests purporting to measure these skills must demonstrate the extent to which test results can be expected to transfer to other contexts.

d. test length; fidelity vs. bandwidth:

Any critical thinking test, whether general or aspect-specific, is limited in the amount of information it can provide. Often decisions in test construction must be made concerning the relative benefits of measuring a broad range of abilities ("bandwidth") vs. sampling fewer abilities more thoroughly, and perhaps with greater accuracy ("fidelity"). Construction of single-aspect tests faces analogous problems in determining the range of conditions under which the single ability will be sampled (see the discussion under "aspect-specific tests", above). The problem is also complicated by several psychometric considerations.

Generally, test reliability increases with test length. Longer tests theoretically also sample particular abilities more intensively. Yet general tests must sample a number of skills in a limited amount of testing time, which often results in subtests being composed of only a few items. These short subtests therefore have low reliabilities, often invalidating them as a basis for individual comparisons.

Subtest length may also reflect the relative contributions of different abilities to the composite test score. The psychometric issues here are beyond the scope of this paper, and the previous statement is somewhat over-simplified. However, a general critical thinking test which devotes 30 of its 50 items to deductive reasoning and the remainder to other abilities, would seem to be primarily a test of deductive reasoning. Similarly, deductive reasoning ability might be expected to contribute more heavily to the composite score on this test than to the composite score on a similar test of equal length with a 15-item deductive reasoning subtest. Differences in subtest length may also reflect differences in the concepts of critical thinking on which the tests were based.

Although many aspect-specific tests do not employ subtests, these tests usually include some variation in the conditions under which particular abilities are sampled (tests of conditional reasoning ability often include emotionally-loaded as well as neutral items, for example). The fidelity vs. bandwidth dilemma also occurs here. Greater item variation may provide a more representative sample of a given skill; test length may also be increased. Reduced variation may result in a more thorough sampling of an ability by a particular type of item, and possibly a test of shorter length. The types of items included, however, may ignore relevant variables which might influence test performance.

Accurate test appraisal must consider the interaction of test length and the fidelity vs. bandwidth problem. High reliability coefficients may be the products of lengthy tests. Such tests, however, may not have sufficient scope for the purposes of some prospective users. Shorter tests may provide a more broadly-based assessment of individual abilities, yet also have lower reliabilities. Neither of these circumstances need obtain, however.

Summary

This review was intended to be of use to individuals concerned with critical thinking testing. Although not comprehensive, it has discussed instruments which represent several distinct approaches to assessment of critical thinking skills, and compared a number of their features. The most common critical thinking instrument is a paper-and-pencil, multiple-choice test scored by the simple number of correct responses. Alternative approaches include individually administered oral interviews and essay tests, which are intended to minimize some of the significant measurement problems associated with more conventional tests. Critical thinking measures may be classified roughly as either general or aspect-specific; this review includes discussions of both types. In general, the measures cited in this review differ widely in such characteristics as test format, target populations, availability, use of subtests, scoring methods, and intended purposes. Such diversity is to be welcomed in a field in which accurate assessment of mental abilities is still an unrealized goal.

References

Ennis, R. H. A concept of critical thinking. Harvard Educational Review, 1962, 32(1), 157-187.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U.S. Government Printing Office, 1975.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978.

APPENDIX:

SUMMARIES AND REVIEWS OF SELECTED CRITICAL THINKING TESTS

ASSUMPTION RECOGNITION INDEX (I AND II)

Author(s): Instructional Objectives Exchange
Year of publication: 1971
Most recent edition: 1971

Description:

The Assumption Recognition Indices are criterion-referenced measures designed to assess examinees' proficiency in recognizing assumptions in a line of reasoning or single statement. The manual defines an assumption as "...a belief for which no proof is offered which underlies an argument or statement" (p. 43).

Index I consists of 5 statements, each followed by a series of proposed assumptions. For each proposed assumption, the examinee must decide whether the speaker must accept the assumption in order to remain "reasonable and consistent." Possible responses are "assumption made" and "assumption not made." Examinees evaluate 25 proposed assumptions in the course of the test. Index II consists of 5 arguments, each followed by four proposed assumptions. Examinees evaluate 20 proposed assumptions in the course of Index II. The manual specifically notes that examinees must understand the meaning of assumption for scores to be meaningful, and cautions test administrators to insure that examinees have this understanding before administering the tests.

A sample problem is provided at the beginning of each test. Scoring is by the 'rights only' method, and only a total test score is computed. A time limit of 10-15 minutes (Index I) or 15-20 minutes (Index II) is suggested for completion. Both Assumption Recognition Indices are intended for use in grades 7-12.

Sample test item (Index I):

IV. STATEMENT: Elect James Riley -- Better City Management.

PROPOSED ASSUMPTIONS:

(9) James Riley is a city management expert.

(10) If elected, Mr. Riley promises to improve city management.

(11) Riley believes the people want better city management.

(12) James Riley is running for the office of mayor.

Reliability and validity information:

No reliability or validity information is provided for either Index, beyond the suggestion given in the manual that items included in each measure were judged appropriate (i.e., congruent with the behavioral objectives) by at least one subject matter expert. However, the meanings of test validity and test reliability remain inadequately specified in the criterion-referenced measurement field. The lack of reliability and validity information in the manual probably reflects this situation, at least in part.

Availability:

Both measures are currently in print and available from the Instructional Objectives Exchange, P.O. Box 24095, Los Angeles, California, 90024.

Comments:

Index I:

This measure (as well as Index II) is noteworthy in that it attempts to assess proficiency in an often-neglected aspect of critical thinking ability — assumption-finding. The test itself is straightforward, and the directions are generally clear. However, the instrument suffers from two problems: (1) assumptions are confused with presuppositions. In several problems, one of the proposed assumptions is actually a presupposition. This confusion is unfortunate because the logical relationship which holds between a statement and its presupposition(s) is not identical to the relation which may hold between a statement and an assumption. Specifically, in order for a statement to be true, the presupposition(s) of that statement must also be true. The truth of the presupposition(s) is logically necessary for the truth of the statement; no such relationship obtains between a statement and an assumption, however. Indeed, we usually speak of assumptions in a line of reasoning, rather than assumptions of a statement. A clarification of the authors' view of the nature of assumptions might prove useful in evaluating the rationale used for test construction, and must surely be a part of any score interpretation for a criterion-referenced measure. (2) A number of the keyed answers are debatable. For example, in the sample test item reproduced here, the scoring key identified proposed assumptions 10 and 11 as being made. The keyed answers are problematic, however. It is not clear that statements 10 and 11 qualify as assumptions under the definition given in the manual (p. 43); although they may be beliefs for which no proof is offered, it is not clear that either statement "underlies" the statement given in III. Further, statements 10 and 11 might be reasonably attributed to the speaker (in the manner of a Grecian conversational implicature), but so might statement 9, which is keyed as an assumption which is not made. The keyed answers seem inconsistent here, and such inconsistency appears

Comments (continued):

in other test items. This reviewer strongly suspects that the authors' failure to specify the meaning of assumption to be used in this test is ultimately responsible for both of the difficulties identified above.

Index II:

This Index actually deals with a different sort of "assumption recognition" than Index I. Although a number of the proposed assumptions are statements which might reasonably be attributed to the speaker of the argument (cf. Index I above), others are gap-fillers in the argument (cf. Ennis, 1969). Although these gap-fillers are not logically necessary for the argument, and are raquired in the sense that with their addition, the argument conforms to a loose deductive model. Unfortunately, however, this additional sense of assumption further complicates interpretation of the authors' intended meaning of the term. Because of this failure to clarify the notion of assumption used in this test, it is not clear whether Index II is intended to measure the same sort of ability as Index I.

In general, the clarification of the authors' notion of assumption seems to be the most pressing problem for both these measures. Such clarification should prove helpful in evaluating the rationale for these instruments, and in appraising the types of interpretations which may be made from them.

Additional information:

Ennis, R. H. Logic in teaching. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1969.

Grice, H. P. Logic and conversation. In D. Davidson & H. Harmon (Eds.), The logic of grammar. Encino, California: Dickenson Publishing, 1975, 64-75.

Instructional Objectives Exchange. Judgment: Deductive logic and assumption recognition. Los Angeles, California: Instructional Objectives Exchange, 1971.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978.

Test citation:

Instructional Objectives Exchange. Assumption Recognition Index (I and II). Los Angeles, California: Instructional Objectives Exchange, 1971.

CLASS REASONING INDEX

Author(s): Instructional Objectives Exchange
Year of publication: 1971
Most recent edition: 1971

Description:

The Class Reasoning Index is a criterion-referenced measure designed to assess proficiency in distinguishing valid from invalid class reasoning arguments. The instrument is based upon 8 principles of class reasoning identified by Ennis and Paulus (1965). These principles include the following:[1]

(1)* particular denial (All A's are B's; Therefore, it is false that some A's are not B's — valid)
(2) full transitivity
(3) full conversion (invalid)
(4)* symmetry of class exclusion (No A's are B's; Therefore, no B's are A's — valid)
(5)* exclusive transitivity (All A's are B's; No B's are C's; Therefore, no A's are C's — valid)
(6) additional relational principles of class logic

The test consists of 32 items, equally divided among the 8 class reasoning principles. The four items testing for each principle consist of two items classified as emotionally-laden, and two items classified as non-emotionally-laden.

Three scores are computed for this test: (1) total test score; (2) score on emotionally-laden items; and (3) score on non-emotionally-laden items. In each case, scoring is done by the 'rights only' method. The manual states that examinees must understand the meanings of valid and invalid if scores on the instrument are to be meaningful.

A special answer sheet is provided with the test. A time limit of 40 minutes is suggested for completion. The Class Reasoning Index is intended for use in grades 7-12.

---

[1] Logical principles discussed in this review are classified by the system proposed by Ennis (1976), whenever possible. An asterisk preceding the name of a principle denotes that the classification is my own.

Sample test item:

6. Given:

No counselors are interesting.

No teachers are interesting.

Then, would this conclusion be valid?

At least some of the counselors are teachers.

Reliability and validity information:

No specific reliability or validity information is provided in the manual, beyond the suggestion that the items included in the measure were judged appropriate (i.e., congruent with the behavioral objectives) by at least one subject matter expert. It should be noted, however, that the meaning of test validity and test reliability has not been adequately specified for criterion-referenced measures. The lack of such information in the manual probably reflects this situation, at least in part.

Availability:

The test is currently in print and available from the Instructional Objectives Exchange, P.O. Box 24095, Los Angeles, California, 90024.

Comments:

The Class Reasoning Index is very similar to the Conditional Reasoning Index (also reviewed here). The format, construction and interpretation of this measure all reflect the criterion-referenced measurement orientation of the Instructional Objectives Exchange personnel.

It seems unnecessary, however, to comment further on this instrument beyond the comments presented for the Conditional Reasoning Index, to which the reader is referred. The Class Reasoning Index suffers from identical difficulties (scope of inferences to be made from test scores about pupil performance; confusion of truth and validity; and lack of criterion-referenced test specifications such as those recommended by Popham (1978)). In the opinion of this reviewer, the measure warrants further development, but is at present incomplete.

Additional information:

Instructional Objectives Exchange. Judgment: Deductive logic and assumption recognition. Los Angeles, California: Instructional Objectives Exchange, 1971.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978.

Test citation:

Instructional Objectives Exchange. Class Reasoning Index. Los Angeles, California: Instructional Objectives Exchange, 1971.

Author(s): Instructional Objectives Exchange
Year of publication: 1971
Most recent edition: 1971

Description:

The Conditional Reasoning Index is a criterion-referenced measure designed to assess proficiency in distinguishing valid from invalid conditional reasoning arguments. The instrument is based upon 12 principles of conditional reasoning identified by Ennis and Paulus (1965). These principles are:[1]

(1) particular inversion (denying the antecedent; an invalid move)
(2) detachment (valid)
(3) particular conversion (affirming the consequent; an invalid move)
(4) particular contraposition (denying the consequent; valid)
(5) full transitivity (valid)
(6) full contraposition (valid)
(7) full conversion (invalid)
(8) forward negative detachment (valid)

and additional 'only if' principles.

The test consists of 48 items, equally divided among the 12 conditional reasoning principles. The four items testing each principle consist of two items classified as emotionally-laden, and two items classified as non-emotionally-laden.

Three scores are computed for this test: (1) total test score; (2) score on emotionally-laden items; and (3) score on non-emotionally-laden items. In each case, scoring is done by the 'rights only' method. The manual states that examinees must understand the meanings of valid and invalid if scores on the instrument are to be meaningful.

A special answer sheet is provided with the test. A time limit of 40 minutes is suggested for completion. The Conditional Reasoning Index is intended for use in grades 7-12.

---

[1] Logical principles discussed in this review are classified by the system proposed by Ennis (1976).

Sample test item:

35. Given:

Mary drinks tea only if it has sugar in it.

Mary is drinking tea.

Then, would this conclusion be valid?

The tea has sugar in it.

Reliability and validity information:

No specific reliability or validity information is provided in the manual, beyond the suggestion that the items included in the measure were judged appropriate (i.e., congruent with the behavioral objectives) by at least one subject matter expert. It should be noted, however, that the meaning of test validity and test reliability has not been adequately specified for criterion-referenced measures. The lack of such information in the manual probably reflects this situation, at least in part.

Availability:

The test is currently in print and available from the Instructional Objectives Exchange, P.O. Box 24095, Los Angeles, California, 90024.

Comments:

The Conditional Reasoning Index represents the only criterion-referenced measure of conditional reasoning ability located for inclusion in this review. As such, it is noteworthy, and in the judgment of this reviewer, the attempt to assess the influence of emotionally-laden material on conditional reasoning ability is also a valuable feature of this instrument. However, the measure suffers from several problems.

Most importantly, the nature of the inferences to be made from test performance is unclear. The manual states that "...scores on [this] measure should be interpreted narrowly, as suggestive of students' ability to make judgments based on logical analysis. They should not be interpreted as comprehensive indices of overall capacity for judgment" (p. 1). In contrast, the scoring instructions presented in the manual note that "The entire measure (48 items) may be administered to obtain a global estimate of the students' ability to reason (p. 11). These statements appear inconsistent, and provide no indication of what sorts of inferences might be justified by scores on this measure. Significantly, no statement is made to the

Comments (continued):

effect that an examinee's score on the test may be taken as an unbiased estimate of the proportion of items from the item population which that examinee would answer correctly. This type of interpretation is usually associated with criterion-referenced tests.

The test items also confuse truth and validity. Strictly speaking, conclusions may be true or false. Arguments are valid or invalid. Test items specifically require examinees to judge the validity of conclusions. This feature of the test is particularly unfortunate since the test authors note that examinees must recognize this distinction for their scores to be meaningful, and the fact that the manual devotes considerable time to discussing the distinction (correctly!).

Finally, this reviewer believes that there is little to distinguish this test from any norm-referenced test of the same ability (with items constructed on the basis of the same conditional reasoning principles). The manual fails to provide the sort of test specifications recommended by Popham (1978). In contrast, the Cornell Conditional Reasoning Test (also reviewed here) probably comes closer to qualifying as a criterion-referenced measure for two reasons: (1) the constraints imposed on the test items are specified more fully; and (2) a greater range of test score interpretations in terms of behaviors is possible (mastery of particular conditional reasoning principles, for example). In the judgment of this reviewer, the Conditional Reasoning Index requires additional development and refinement if it is to be used as a (criterion-referenced) measure of conditional reasoning ability.

Additional information:

Instructional Objectives Exchange. Judgment: Deductive logic and assumption recognition. Los Angeles, California: Instructional Objectives Exchange, 1971.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978.

Test citation:

Instructional Objectives Exchange. Conditional Reasoning Index. Los Angeles, California: Instructional Objectives Exchange, 1971.

CORNELL CLASS REASONING TEST, FORM X

Author(s): Robert H. Ennis, et al.
Year of publication: 1964
Most recent edition: 1964

Description:

This test assesses mastery of eight class logic principles and follows a format similar to that of the Cornell Conditional Reasoning Test. The test consists of 72 scored items and 6 sample items which may be used for practice and are based on logical principles not tested for in the instrument. The 72 items are divided into 12-item groups. Eight of these item groups test for mastery of the basic principles. The remaining four item groups require the examinee to apply the basic principles in combination. In addition, the content type of the test items is varied. In each item group, four items have content classified as concrete familiar, one has content classified as symbolic, and the remaining item has suggestive content, in which the examinee is asked to assume the truth of obviously false premises.

The test was designed for and used in grades 4-12, and contains clear, thorough (but not long-winded) instructions. No time limit is suggested; the instructions explicitly state that this is not a speed test. Students are cautioned to avoid wild guessing, and to omit items for which they do not know the answer. A correction formula (rights minus 1/2 wrongs) is employed.

Sample test item:

Suppose you know that

All the members of the school band have been in Boston.

No one in Frank's class has been in Boston.

Then would this be true?

At least some members of the school band are in Frank's class.

A. YES

B. NO

C. MAYBE

The following interpretation of the responses is provided at the top of each page:

YES	It must be true.
NO	It can't be true.
MAYBE	It may be true or it may not be true. You weren't told enough to be certain whether it is "YES" or "NO".

Reliability and validity information:

Test-retest reliability estimates for both part and total test scores were computed on the basis of testing 329 students in grades 4, 6, 8, 10, and 12 in New York state. Approximately ten weeks elapsed between test administrations. Reliability estimates for total test score ranged from .66 (grade 8) to .88 (grade 6) with a mean $r = .83$. Reliability of item subgroup scores ranged from .26 [grade 6, combination of particular denial (invalid) and full contraposition (valid) principles] to .77 [grade 12], particular denial (invalid)]. Across grade levels, mean reliability estimates for each principle varied between .69 [full conversion (invalid)] and .36 [combination of particular denial (invalid) and full contraposition (valid) principles].[1]

Content validity of the test was established by a careful examination of the field of logic and specific examples of common logical reasoning, including newspaper editorials, U.S. Supreme Court opinions and a handbook for auto mechanics. The test was also taken and judged to be valid by Keith Donellan, an expert in the field of logic. Construct validity of the test was partly supported when significant hypothesized correlations between total test score and I.Q. were obtained for each grade level and for all levels combined ($p < .05$). Total test score was also uncorrelated with sex for all grades combined, and correlated significantly with chronological age when correlations for each grade level were averaged ($r = .68$; $p < .05$). Validity estimates were based on the first test administration. Although the report which serves as the test manual does not explicitly provide norms, means and standard deviations are reported for total test score on both pretest and posttest for all groups. Additional information on the research subjects (chronological age, I.Q., socio-economic status) is also provided when available.

Availability:

The test is currently in print and is available on request from the Illinois Critical Thinking Project, Bureau of Educational Research, 188 Education Building, University of Illinois, Urbana, Illinois, 61801. A comprehensive report describing the project for which this test and one other were developed serves as the test manual (see below). The report is available on ERIC (# ED 003 818) and in libraries that received Cooperative Research Reports.

Comments:

The Cornell Class Reasoning Test, Form X appears to be a reasonably valid and reliable measure of mastery of eight principles of class logic.

---

[1] Whenever possible, the logical principles discussed in this review are classified by the system suggested by Ennis, (1976). 'Particular denial' denotes the following class logic principle:

All A's are B's.

Therefore, at least some A's are not B's. (invalid)

Comments (continued):

Test directions are clear and straightforward, and the vocabulary is not too difficult for the recommended use in grades 4-12. Although items testing for mastery of the easier principles appear earlier in the test, a careful distribution of individual items was employed in order to avoid any patterning of responses. Value-laden and ambiguous language is also minimized.

For the purposes of the study in which this instrument was used, the researchers operationally defined mastery of a logical principle as at least five correct responses out of the six questions for a given principle. The design of the test also allows discrimination between students who may have mastered a particular class reasoning principle for one content type and those who have not. A student might correctly answer the four concrete familiar test items for several principles but consistently miss the other two. According to the criterion for mastery, he would not have mastered the principles. However, this response pattern might also suggest that this student has mastered some class logic principles for items with concrete familiar content. Because of the limited number of items per group, however, the test could probably be used only to identify students who might have mastered class logic principles for concrete familiar content, since four of the six items in each group are of this type.

Form X is currently the only form of the Cornell Class Reasoning Test which has been developed.

Additional information:

Ennis, R. H. An alternative to Piaget's conceptualization of logical competence. Child Development, 1976, 17, 903-919.

Ennis, R. H., & Macmillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Ennis, R. H. & Paulus, D. H. Critical thinking readiness in grades 1-12, (Phase I, deductive logic in adolescence). (U.S. Office of Education Publication No. 1680). Washington, D.C.: U.S. Government Printing Office, 1965. (ERIC Document Reproduction Service No. ED 003 818)

Test citation:

Ennis, R. H., Gardiner, W. L., Morrow, R., Paulus, D., & Ringel, L. The Cornell Class Reasoning Test, Form X. Ithaca, New York: Cornell Critical Thinking Project, 1964.

Author(s): Robert H. Ennis, et al.
Year of publication: 1964
Most recent edition: 1964

Description:

The Cornell Conditional Reasoning Test, Form X was designed to assess mastery of ten conditional reasoning principles, and is similar in format to the Cornell Class Reasoning Test. The test consists of 72 scored items and 6 sample items which may be used for practice and are based on principles not tested for in the instrument. The 72 scored items are divided into 12 six-item groups. Ten of these groups test for mastery of the basic principles, while the remaining two groups require examinees to apply the basic principles in combination. In addition, the content type of the test items is varied. In each item group, four items have content classified as concrete familiar, one has content classified as symbolic, and the remaining item has suggestive content, in which the examinee is asked to assume the truth of obviously false premises. Within each item group, and when allowed by the argument, one concrete familiar item presents a conclusion which is a negation of what would follow necessarily from the premises. Another concrete familiar item presents one premise as a negation, and for those items having at least two premises, the usual order of presentation is altered.

The test was designed for and used in grades 5-11, and contains clear, thorough (but not long-winded) instructions. No time limit is suggested; the instructions explicitly state that this is not a speed test. Students are cautioned to avoid wild guessing, and to omit items for which they do not know the answer. A correction formula (rights minus 1/2 wrongs) is employed.

Sample test item:

Suppose you know that

If the chair is green, then the table is black. A. YES

Then would this be true? B. NO

If the table is black, then the chair is green. C. MAYBE

The following interpretation of the responses is provided at the top of each page:

YES	It must be true
NO	It can't be true
MAYBE	It may be true or it may not be true. You weren't told enough to be certain whether it is "YES" or "NO".

Reliability and validity information:

Test-retest reliability estimates for both part and total test scores were computed on the basis of testing 266 students in grades 5, 7, 9, and 11 in New York state. Approximately ten weeks elapsed between test administrations. Reliability estimates for total test score ranged from .65 (grade 7) to .80 (grade 11) with a mean r = .75. Reliability of item subgroup scores ranged from .27 [grade 7, denial of particular contraposition (invalid)] to .70 [grade 9, combination of detachment (valid) and full transitivity (valid)]. Across grade levels, mean reliability estimates varied between .41 [denial of particular contraposition (invalid)] and .60 [combination of detachment (valid) and full transitivity (valid)].[1]

Content validity of the test was established by a careful examination of the field of logic and specific examples of common logical reasoning, including newspaper editorials, U.S. Supreme Court opinions, and a handbook for auto mechanics. Construct validity of the test was partly supported when significant hypothesized correlations between total test score and I.Q. were obtained for each grade level and for all levels combined ($p < .05$). Total test score was also uncorrelated with sex for all grades combined, and correlated significantly with chronological age when correlations for each grade level were averaged ($r = .58$; $p < .05$). Validity estimates were based on the first test administration. Although the report which serves as the test manual does not explicitly provide norms, means and standard deviations are reported for total test score on both pretest and posttest for all groups. Additional information on the research subjects (chronological age, I.Q., socio-economic status) is also provided when available.

Availability:

The test is currently in print and is available on request from the Illinois Critical Thinking Project, Bureau of Educational Research, 188 Education Building, University of Illinois, Urbana, Illinois, 61801. A comprehensive report describing the project for which this test and one other were developed serves as the test manual (see below). The report is available on ERIC (# ED 003 818) and in libraries that received Cooperative Research Reports.

---

[1] Whenever possible, the logical principles discussed in this review are classified by the system suggested by Ennis, (1976). 'Denial of particular contraposition' denotes the following conditional reasoning principle:

If P, then Q.
Not Q.
Therefore, P. (invalid)

Comments:

The Cornell Conditional Reasoning Test, Form X appears to be a reasonably valid and reliable measure of mastery of ten principles of conditional logic. Test directions are clear and straightforward, and the vocabulary is not too difficult for the recommended use in grades 5-12. Although items testing for mastery of the easier principles appear earlier in the test, a careful distribution of individual items was employed in order to avoid any patterning of responses. Value-laden and ambiguous language is also minimized.

For the purposes of the study in which this instrument was used, the researchers operationally defined mastery of a logical principle as at least five correct responses out of the six questions for a given principle. The design of the test also allows discrimination between students who may have mastered a particular conditional reasoning principle for one content type and those who have not. A student might correctly answer the four concrete familiar test items for several principles but consistently miss the other two. According to the criterion for mastery, he would not have mastered the principles. However, this response pattern might also suggest that this student has mastered some conditional logic principles for items with concrete familiar content. Because of the limited number of items per group, however, the test could probably be used only to identify students who might have mastered conditional reasoning principles for concrete familiar content, since four of the six items in each group are of this type. It should also be noted that an item analysis performed on the test suggested that some conditional reasoning principles may be beyond the grasp of younger students. Although the overall mean difficulty index for grade 5 students was 47.5, principles 3 and 7 had mean difficulty indices of 17.7 and 14.1 respectively, which indicates that only a small percentage of students were able to answer correctly the items testing for mastery of these principles.

Form X is currently the only form of the Cornell Conditional Reasoning Test which has been developed.

Additional information:

Ennis, R. H. An alternative to Piaget's conceptualization of logical competence. Child Development, 1976, 17, 903-919.

Ennis, R. H., & Macmillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Ennis, R. H., & Paulus, D. H. Critical thinking readiness in grades 1-12. (Phase I, deductive logic in adolescence). (U.S. Office of Education Publication No. 1680). Washington, D.C.: U.S. Government Printing Office, 1965. (ERIC Document Reproduction Service No. ED 003 818)

Test citation:

Ennis, R. H., Gardiner, W. H., Guzzetta, J., Morrow, R., Paulus, D., & Ringel, L. The Cornell Conditional Reasoning Test, Form X. Ithaca, New York: Cornell Critical Thinking Project, 1964.

CORNELL CRITICAL THINKING TEST, LEVEL X

Author(s): Robert H. Ennis and Jason Millman
Year of publication: 1961
Most recent edition: 1971

Description:

The Cornell Critical Thinking Test, Level X, was one of two tests developed by Ennis and Millman to assess the degree of mastery of a limited number of aspects of critical thinking ability. This test is designed to assess mastery of the following four aspects of critical thinking ability: (1) inductive reasoning -- evaluation of evidence and its direction of support for a given hypothesis; inference to best explanation; (2) judging reliability of observation statements and inferences made from them; (3) deductive reasoning ability -- judging what follows necessarily from given premises; and (4) assumption-finding. Each aspect is tested in a separate section. The test consists of 71 scored items and 5 sample items which are used to familiarize examinees with the form of the questions in each part of the test. The 71 scored items are divided as follows; induction, 23 items; judging reliability, 24 items; deductive reasoning, 9 items; and assumption-finding, 15 items. Although the test is conveniently divided into four sections, the authors discourage the use of part scores on the basis that such scores are likely to have "borderline reliability" (manual, p. 8).

The test is written in the form of a story about a group of explorers who land on an unfamiliar planet and must deal with the problems they encounter there. This format seems both interesting and effective for sampling the abilities the test is designed to measure. Clear and straightforward directions are provided. Although the authors provide no explicit statement of the age groups for which the test was designed, the manual notes that the test "...is taken and enjoyed by students from the junior high (about age 14) on up" (p. 2), which suggests the test should be appropriate for students in grades 7-12, and possibly those in the first two years of college as well. A time limit of 50 minutes is suggested, although the authors note that the limit may be extended if comparisons with the norms presented in the manual are unimportant. Students are cautioned to avoid wild guessing, and to omit items for which they do not know the answer. A correction formula (rights minus 1/2 wrongs) is employed (however, see the information on norms below).

A manual is provided for the test, and contains discussions of the concept of critical thinking on which the test is based; the aspects of critical thinking ability tested for in the four test sections, test administration and scoring; norms; and reliability and validity estimates. The manual is currently under revision.

Sample test item (assumption-finding):

"The shorter of the two people wearing green hats is a female. I know because I saw her long hair when she removed her hat." Which is taken for granted?

A. All females have long hair.

B. Only females have long hair.

C. A person wearing a green hat is likely to be a female.

Reliability and validity information:

Spearman-Brown and Kuder-Richardson 20 reliability estimates for total test score were computed on four groups of students ranging from grade 7 to senior high level, with a total N = 2767. Reliability estimates ranged from .77 (Kuder-Richardson 20; senior high school students) to .87 (Spearman-Brown; averaged over grades 7 and 8). Total test scores were computed by the 'rights only' formula for one group and the rights minus 1/2 wrongs formula for the remaining three groups.

The manual presents no arguments for the content validity of the test, beyond the information given in discussions of the concept of critical thinking on which the test was based. Construct validity of the test was supported in part by obtained correlations between total test score on the Cornell Critical Thinking Test, Level X, and performance on other criterion measures. Correlations are presented for three groups of students: (1) grade 7 and 8 students in an upstate New York school (N = 110); (2) grade 8 students in eleven high schools in upstate New York (N = 1126); (3) education students at a pacific coast state university (N = 127 or 187, depending upon correlation computed). Although the manual states that the obtained correlations are "unsurprising" (p. 12), several correlations appear especially relevant to the construct validity argument. These correlations include r = .55 (California Test of Mental Maturity -- Language; group 1); r = .45 (California Test of Mental Maturity -- Total I.Q.; group 1); r = .53 (New York State Junior High Survey Test -- Vocabulary; group 1); r = .52 (Scholastic Aptitude Test -- Verbal; group 3); and r = .52 (Scholastic Aptitude Test -- Total; group 3). The authors neglect to report the scoring method by which total test scores on the Cornell Critical Thinking Test, Level X were computed for groups 1 and 2.

Norms are provided for various groups of students, ranging from grade levels 7 and 8 to senior high school. However, the authors were able to report only the mean score and standard deviation for the group of 187 education students at a pacific coast state university. Mean scores and standard deviations are also provided for the other groups tested. When available, norms, means and standard deviations are provided for the 'rights only' scoring formula as well as the rights minus 1/2 wrongs scoring formula. The manual states that these norms were secured "...in the 1960's" (p. 8).

Availability:

The test is currently in print and is available on request from the Illinois Critical Thinking Project, Bureau of Educational Research, 188 Education Building, University of Illinois, Urbana, Illinois, 61801. A specimen set (one test, manual with answer key; answer sheet) is also available at the above address.

Comments:

The Cornell Critical Thinking Test, Level X appears to be a reasonably valid and reliable measure of proficiency in some aspects of critical thinking ability. Test directions are clear, and the sample items preceding each test section seem adequate to insure that examinees understand what is expected of them throughout the test. For the most part, the language used in the test is unambiguous and appropriate for the grade levels for which the test is recommended.

The major difficulty with this test is in the first 23 items (induction). Although good arguments can be made for the keyed answers, equally strong arguments might be made for other responses. Because of this fact, students who choose an incorrect answer for justifiable reasons may be penalized. It is important to note, however, that this problem is one common to tests of inductive reasoning ability, where inferences must be drawn on the basis of the relatively little information given in the problem. In addition, the validity of examinee responses to the induction items could be affected by a failure to distinguish evidence which might *support* a hypothesis from evidence which might *establish* the same hypothesis. The test directions specifically caution against this error, however.

The test manual is fairly complete, although more current norms and reliability and validity information would be a welcome addition. Such information is forthcoming in the revised edition of the test manual. No intercorrelations among subtests are reported, although this omission perhaps reflects the contention of the authors that individual subtests are likely to be unreliable (hence the recommendation against the use of part scores).

The Cornell Critical Thinking Test, Level X is currently one of the more widely-used critical thinking tests on the market. Level Z of the test has also been developed and is somewhat more difficult. The Level Z version is also a substantially different instrument from the Level X version.

Additional information:

Ennis, R. H. A concept of critical thinking. *Harvard Educational Review*, 1962, *32*(1), 157-187.

Ennis, R. H. *Logic in teaching*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1969.

Additional information (continued):

Ennis, R. H., & Macmillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Follman, J. A factor analytic study of three critical thinking tests, one english test, and one logical reasoning test. Doctoral dissertation, Indiana University, 1969.

Follman, J. Correlational and factor analysis of critical thinking, logical reasoning, and english total test scores. Florida Journal of Educational Research, 1970, 12(1), 91-94.

Follman, J. Factor analysis of three critical thinking tests, one logical reasoning test, and one english test. Yearbook of the National Reading Conference, 1969, 18, 154-160.

Follman, J., Brown, L., & Burg, E. Factor analysis of critical thinking, logical reasoning and english subtests. Journal of Experimental Education, 1970, 39(4), 11-16.

Follman, J., Brown, L., & Burg, E. Canonical correlation of scholastic aptitude and critical thinking. Psychology, 1969, 6(3), 3-6.

Follman, J., Miller, W., & Hernandez, D. Factor analysis of achievement, scholastic aptitude and critical thinking subtests. Journal of Experimental Education, 1969, 38(1), 49-53.

Test citation:

Ennis, R. H., & Millman, J. Cornell Critical Thinking Test, Level X. Urbana, Illinois: Illinois Critical Thinking Project, University of Illinois, 1961.

CORNELL CRITICAL THINKING TEST, LEVEL Z

Author(s): Robert H. Ennis and Jason Millman
Year of publication: 1961
Most recent edition: 1971

Description:

The Cornell Critical Thinking Test, Level Z was designed to assess mastery of a limited number of critical thinking skills. The test consists of 52 items and is divided into seven sections. The aspects of critical thinking ability tested for in each section are as follows: (1) Section 1 (items 1-10): deductive reasoning; determining whether a statement follows from given premises. Emotionally-loaded material is deliberately introduced. (2) Section 2 (items 11-21): identifying faulty reasoning; recognition of informal fallacies such as circularity, oversimplification of alternatives and non-supporting emotive language. (3) Section 3 (items 22-25): judging reliability of statements. (4) Section 4 (items 26-38): evaluation of evidence and its direction of support for a given hypothesis. (5) Section 5 (items 39-42): choosing useful hypothesis-testing predictions. (6) Section 6 (items 43-46): assumption-finding; determining what a speaker probably had in mind by a particular usage of a term. (7) Section 7 (items 47-52): assumption-finding (gap-filling); identification of the unspoken premise in a speaker's argument. Although the test is divided into seven sections, the authors discourage the use of part scores on the basis that such scores are likely to have "borderline reliability" (manual, p. 8).

The test contains clear directions, but no sample problems for examinees to work. In general, the Level Z version of the Cornell Critical Thinking Test is somewhat more difficult than the Level X version, and follows a different format. Although the authors provide no explicit statement of the age groups for which the test was intended, they do say that Level Z "...is a bit formidable for the average secondary student...[and]...is best adapted to people in higher education, but high ability secondary students can cope with it" (manual, p. 2). In addition, reliability and validity information and norms are reported for students in grades 10 to graduate school, which also gives some idea of the range of the test's applicability. A time limit of 50 minutes is suggested, although the authors note that more advanced students tend to spend more time on the test, and the limit may be extended if comparisons with the norms presented in the manual are unimportant. Examinees are cautioned to avoid wild guessing, but are not told whether or not to omit items for which they do not know the answer. A correction formula (rights minus 1/2 wrongs) is employed (however, see the information on norms below).

A manual is provided for the test, and contains discussions of the concept of critical thinking on which the test is based; the aspects of critical thinking ability tested for in each test section; test administration and scoring; norms; and reliability and validity estimates. The manual is currently under revision.

Sample test item (assumption-finding; section 6):

"It certainly is a stock car," says Jim. "It has an ordinary engine that hasn't been changed since it came off the assembly line. That alone makes it a stock car and that's all that matters."

Of the following, what is the best way to state Jim's notion of a stock car?

A. A stock car is an automobile that is for the most part made of standard parts put out by automobile manufacturers, but which might have the fenders missing and special type bumbers.

B. A stock car is an automobile with a standard engine.

C. A stock car is where the engine is standard.

Reliability and validity information:

Spearman-Brown, Kuder-Richardson 20 and Kuder-Richardson 21 reliability estimates for total test score were computed for nine groups of students ranging from grade 10 students to graduate students with a total N = 1228. Reliability estimates varied between .55 (Spearman-Brown; undergraduate students) and .77 (Kuder-Richardson 21; community college students). Total test score was computed by the 'rights only' formula for four groups and the rights minus 1/2 wrongs formula for the remaining five groups. The authors state that the Kuder-Richardson 21 reliability estimates may be inflated when total test score was computed with the rights minus 1/2 wrongs formula.

The manual presents no arguments for the content validity of the test, beyond the information given in discussions of the concept of critical thinking on which the test is based. Construct validity of the test was supported in part by obtained correlations between total test score on the Cornell Critical Thinking Test, Level Z and performance on other criterion measures. Correlations are presented for four groups of students: (1) undergraduate students in junior-level educational psychology at a midwestern state university (N = 190); (2) grade 10 and 11 students from a northeastern suburban high school with a high proportion of college-bound students (N = 35); (3) undergraduates in an introductory biology course at an Ivy League university (N = 109; validity estimates for subsamples of this group are also reported); and (4) fifth-year teaching interns at the same institution (N = 22). Although the manual states that the obtained correlations are "unsurprising" (p. 12), several correlations appear especially relevant to the construct validity argument. These correlations include r = .44 (A Test of Critical Thinking, Form G; group 1); r = .48 (Watson-Glaser Critical Thinking Appraisal, Form Zm; group 1); r = .58 (School and College Ability Test; group 2 grade 10 students); r = .71 (School and College Ability Test; group 2 grade 11 students); r = .51

Reliability and validity information (continued):

(Scholastic Aptitude Test -- Math; group 3); and r = .44 (Test of Understanding Science; group 3). Unfortunately, the authors neglect to report the scoring method by which total test score on the Cornell Critical Thinking Test, Level Z was computed for group 4.[1]

Norms are provided for various groups of students, ranging from undergraduate to graduate level. Mean scores and standard deviations are also reported for the groups tested. When available, norms, means and standard deviations are provided for the 'rights only' scoring method as well as for the rights minus 1/2 wrongs scoring formula. The manual states that these norms were secured "...in the 1960's" (p. 8).

Availability:

The test is currently in print and is available on request from the Illinois Critical Thinking Project, Bureau of Educational Research, 188 Education Building, University of Illinois, Urbana, Illinois, 61801. A specimen set (one test, manual with answer key; answer sheet) is also available at the above address.

Comments:

The Cornell Critical Thinking Test, Level Z appears to be a reasonably valid and reliable measure of proficiency in some aspects of critical thinking ability. Test directions are clear, although the addition of sample items preceding each section would be welcome. Perhaps the authors felt that because the test is aimed at more advanced students, sample items were not required. The language of the test includes some technical terms (such as "follows necessarily") but these are defined in the directions for the part of the test in which they appear. For the most part, the language used is unambiguous and appropriate for the grade levels for which the test is recommended.

The section of the test dealing with identifying fallacious reasoning was the most problematic for this reviewer. The items in this section require examinees to judge which of three suggested reasons best explains why the reasoning presented in a particular passage is faulty. This reviewer felt that the best explanation for the faulty reasoning sometimes did not appear among the alternatives. However, the test directions instruct examinees to "pick the one best reason why some of this thinking is faulty", and

---

[1] Validity estimates for group 4 are conspicuously absent from this review. The reason for this is that their relevance for the construct validity argument is unclear. This issue is discussed further below ("Comments").

Comments (continued):

a "none of the above" alternative is not included, which suggests that only the alternatives provided should be considered. This reviewer also disagreed with some of the keyed answers in this section.

Test sections 3 (judging reliability of statements and evaluation of warrant), 5 (choosing useful hypothesis-testing predictions) and 6 (assumption-finding; definitions) each contain only four items. It seems likely that such a limited number of items is insufficient to provide an adequate sample of the ability being tested. Although the authors discourage the use of part scores and note that the third test section serves "...as a rough check on [the] assumption [that students for whom the test was designed should have already mastered the ability to judge the reliability of statements]" (manual, p. 3), no rationale is presented for including only a limited number of items in sections 5 and 6. It is not clear how scores on these three test sections should be interpreted. (In addition, the same argument might be made against section 7 -- assumption-finding; gap-filling -- which contains only six items).

The test manual is fairly complete, although more current norms and reliability and validity information would be a welcome addition. Such information is forthcoming in the revised edition of the test manual. It would also be interesting to see the extent to which performance on the Cornell Critical Thinking Test, Level Z correlated with performance on the Level X version. No intercorrelations among subtests are reported, although this omission perhaps reflects the contention of the authors that individual subtests are likely to be unreliable (hence the recommendation against the use of part scores).

The construct validity argument is clouded somewhat by the validity coefficients for group 4 (teaching interns). All but one of the correlations reported were obtained for performance on subtests of the Allport-Vernon-Lindzey Study of Values, which, as the authors note, is an ipsative test (i.e., scores on any one scale are determined by scores on the remaining scales). Ennis and Millman state that "...correlations between these scale scores and an normative measure (such as the Cornell Critical Thinking Test) are not expected to be particularly high" (manual, p. 14). The low obtained correlations, then, may be statistical artifacts and therefore unrepresentative of the true strength of association between critical thinking ability as measured by the Cornell Critical Thinking Test, Level Z and the traits measured by the various scales of the Allport-Vernon-Lindzey Study of Values. It is also difficult to know whether the obtained correlation of $r = -.53$ between total test score on the Cornell Critical Thinking Test, Level Z and performance on the Minnesota Teacher Attitude Inventory for this group says more about teacher attitudes or more about the validity of the test. In the absence of any theoretical system which might lead to predictions about correlations between performance on this test and other criterion measures, evaluation of the construct validity of the test remains difficult.

Comments (continued):

One must read the test and judge for oneself, given the limited information in the current manual.

Additional information:

Ennis, R. H. A concept of critical thinking. Harvard Educational Review, 1962, 32(1), 157-187.

Ennis, R. H. Logic in teaching. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1969.

Test citation:

Ennis, R. H., & Millman, J. Cornell Critical Thinking Test, Level Z. Urbana, Illinois: Illinois Critical Thinking Project, University of Illinois, 1961.

Author(s): DeWitt E. Sell
Year of publication: 1952
Most recent edition: 1952

Description:

The manual for the Evaluation Aptitude Test defines evaluation aptitude as "...the capacity to appraise data accurately; to draw correct inferences from given premises; to think 'straight' or without emotional bias" (p. 1). The test is intended to measure (1) "...a subject's composite aptitude to reason deductively" and (2) "...the degree to which the emotional bias factor is influencing a subject's deductive reasoning" (manual, p. 1). Examinees are presented with 36 syllogisms in the course of the test, and must decide if the conclusion presented is sound (follows necessarily), unsound (does not follow necessarily) or if they are uncertain of whether the conclusion is sound or unsound. The instrument is divided into two parts, each containing 18 class-reasoning syllogisms. According to the manual, the items in Part A contain material intended to be emotionally neutral. In addition, Sell states that the minor term of the syllogisms in Part A is not factually related to either the major or middle terms of the syllogisms. Sell uses 'twangs' as the minor term in every syllogism in Part A; since this is a nonsense word, it cannot be factually related to either 'candies' or 'confections', which serve as the major and middle terms of the syllogisms in Part A. Of the 18 items in this part, 12 are invalid syllogisms and 6 are valid. According to the test manual, this proportion was selected for two reasons: (1) fallacious syllogistic moods greatly outnumber valid ones; and (2) examinees who assume an equal number of valid and invalid syllogisms will be disrupted. Sell also claims that all the invalid syllogistic forms are represented except the fallacy of four terms (ambiguous middle).

Part B is identical to Part A except that the syllogisms employ emotionally-loaded material in the premises and conclusion. Test items of Part B correspond exactly to those in Part A in logical form; for example, item 1 of Part A and item 19 of Part B are syllogisms identical in logical form and differ only in the emotional content of the premises and conclusion. This relationship is maintained throughout the test. An attempt was also made to construct 9 syllogisms whose fallacious conclusions would likely be affirmed by liberals or whose valid conclusions would likely be denied by liberals. The remaining 9 syllogisms of this part are intended to contain fallacious conclusions likely to be affirmed by conservatives and valid conclusions likely to be denied by conservatives. The intent of this device is to make a high score on Part B indicative of ability to evaluate arguments on purely logical grounds, and to ignore the emotional content of the premises and conclusion.

Five scores are computed for this test: (1) score on Part A; (2) score on Part B; (3) total test score; (4) index of discrepancy; and (5) index of indecision. The part scores are computed by a twice rights minus wrongs formula; this is equivalent to a traditional correction for guessing. Items

Description (continued):

marked 'O' (uncertain) are ignored in obtaining part scores. Total test score is computed as the sum of scores on Part A and Part B. The index of discrepancy is obtained by comparing responses to Part B items with the responses to the logically equivalent Part A items. The total of Part B responses which differ from their Part A logical counterparts is divided by 18 to yield the index of discrepancy. The index of indecision is calculated as the total number of 'uncertain' responses (in both parts of the test) divided by 36.

No recommendations concerning the age groups for which the test might be appropriate are made in the test manual. However, norms for Part A, Part B and total test scores are presented for high school seniors and college seniors. Index of discrepancy and index of indecision norms are given for college seniors. The manual also states that "...the [Evaluation Aptitude Test] as an aptitude test is meaningless for any person who has had training in deductive logic" (p. 2), and that the test is designed for individuals who have at least completed high school (p. 3).

A time limit of 50 minutes is specified for test completion. Sell states that the majority of examinees complete the test within the time limit. The test is intended to be a power test. Test directions are generally clear, and appear only at the beginning of the test. The test itself consists of a single sheet containing the 36 syllogisms and answer marking spaces on one side, and test directions and spaces for summary information (such as part and total test scores) on the other. Test directions explicitly define 'some' as meaning 'at least some' (i.e., some and possibly all). Examinees are not informed that a correction for guessing will be imposed, and are not explicitly cautioned to avoid guessing.

Sample test item:

[The following directions appear at the beginning of the test]:

This test contains 36 items which you are to evaluate. Each item has two propositions and a conclusion based upon these propositions. Your task is to judge whether the conclusion NECESSARILY FOLLOWS from the statements AS GIVEN in each case. If you are CERTAIN the conclusion FOLLOWS, fill in the pair of lines under S (Sound) in the margin to the right of the item. If you are CERTAIN the conclusion DOES NOT NECESSARILY FOLLOW, fill in the pair of lines under U (Unsound) in the margin to the right of the item. If you are UNCERTAIN about the soundness of any conclusion, you should indicate this by filling in the pair of lines under O to the right of that item.

Sample test item (continued):

The word "some" as used in these items means "some and possibly all," i.e., "at LEAST some." There are two parts to the test. When you have completed Part A, go right on to Part B. Follow these same directions in both parts. You will be allowed 50 minutes for the test. Most individuals finish before that length of time, so do not feel hurried.

[capitalization in original]

Part A	Part B
No candies are confections.	No socially mature individuals are political radicals.
All confections are twangs.	All political radicals are Liberals.
Hence, some twangs are not candies.	Hence, some Liberals are not socially mature individuals.
S U O ☐ ☐ ☐	S U O ☐ ☐ ☐

Reliability and validity information:

The manual presents considerable validity and reliability information for this test. For the purposes of this review, the information given in the test manual is categorized below.

item validity: For each part of the test, the same group of high school seniors (N = 165) was divided into quartiles on the basis of part scores. Discrimination values for each item were calculated by subtracting the percentage of students in the fourth quartile missing a given item from the percentage of first quartile students responding incorrectly to the same item. A separate analysis was performed for each part of the test. The results of these analyses showed that in all cases, at least one of the logically equivalent items comprising an item-pair discriminated at or beyond the 5% level of confidence. The manual states that the 5% level of confidence corresponds to a difference of 21 percentage points.

Part B validity: A sample of 43 students from a class in General Psychology and Mental Hygiene from Huntingdon College in Alabama were administered the Evaluation Aptitude Test. Subsequently, these subjects were asked to indicate their agreement or disagreement with the conclusions of the arguments presented in Part B. The conclusions were presented in the form of a separately-administered attitude survey. Part B scores were computed as described above. Scores on the aptitude survey were calculated as the total number of items for which 'agree' responses matched 'sound' responses

Reliability and validity information (continued):

on Part B of the test, and 'disagree' responses matched 'unsound' responses on Part B. Scores on the attitude survey were then correlated with Part B scores, yielding a coefficient of r = -.55. Sell concludes that this negative correlation "...validates the proposition that to attain a high score on Part B the examinee must not only have considerable aptitude to reason deductively but must have rational control of his emotional biases" (manual, p. 2).

In another study of 60 college students, a coefficient of contingency of .41 was found when the direction of errors on Part B was correlated with the ideology of each subject as measured by a liberal-conservative attitude scale. Sell neglects to report whether the attitude scale was the same instrument used in the study cited above.

An attitude survey consisting of the 18 Part B conclusions was also administered to 40 faculty members at Huntingdon College. Although no data is presented in the manual, Sell claims that the results of the study show that "...deductive reasoning aptitude plus rational control of emotional bias is demonstrated on Part B only when scores are higher than 9 and to the extent that scores are higher than 9..." (manual, p. 2).

validity of total test score: A sample of 44 college students were administered the Evaluation Aptitude Test as a pretest and again as a posttest following a one-semester course in logic. Sell reports the following results: (1) the mean pretest score on Part B was significantly greater than the mean pretest score on Part A; (2) posttest mean scores on each part did not differ significantly; (3) the pretest mean index of discrepancy was found to be .33. The corresponding posttest mean was found to be .12. Sell concludes that the remaining discrepancy evidenced in the posttest administration is primarily due to the continued operation of emotional bias (manual, p. 2).

A further study is reported in which total test and part scores for 64 business agents in five AFL Building Trades local unions were correlated with five criteria of job success. These criteria were: (1) leadership rank (as determined by a panel of 4 judges); (2) votes received in union elections; (3) ability to recruit and organize (as determined from union records); (4) ability to settle grievances and disputes (as determined by a panel of 5 union officials); and (5) ability to enforce rules and regulations. Intercorrelations ranged from r = -.33 (index of indecision and ability to settle grievences) to r = .27 (total test score and ability to enforce rules).

Other correlations of total test score and scores for Part A and Part B are also reported for a sample of Huntingdon College seniors (N = from 60 to 95, depending upon correlation computed). These scores on the Evaluation Aptitude Test were correlated with grade point averages and ACE test scores. Correlations varied between r = .39 (Part A score and grade point average) and r = .54 (total test score and ACE score).

Reliability and validity information (continued):

total test reliability: A sample of 64 undergraduates in Sell's psychology classes was used to compute a test-retest coefficient of reliability. The period between test administrations was "over two months" (manual, p. 3). The correlation between the two sets of scores was found to be r = .88. Sell also reports the index of reliability for this test to be r = .94. Another reliability coefficient obtained by the method of "rational equivalence" is reported as r = .89. Sell provides no details concerning how this coefficient was obtained.

A split-half (odd-even) reliability coefficient was also computed on the basis of a sample of high school seniors (N = 165). The test manual reports the coefficient as r = .71, but notes that the assumption of an ordered difficulty of test items (easiest to most difficult) was not met, since item analysis revealed the odd-numbered items to be significantly more difficult than even-numbered items (no level of significance reported). Sell therefore concludes that the obtained split-half reliability coefficient "...does not adequately reflect the self-consistency of the test" (manual, p. 3).

part score reliability: The reliability for the index of discrepancy score for a sample of 64 college students is reported as r = .45. The method used for calculation of this coefficient is not stated.

norms: Norms for both high school and college seniors are presented for total test score, Part A score and Part B score. Median scores, mean scores, standard deviations and standard errors are also reported for the groups specified above on the same scores. In addition, college senior norms for index of discrepancy and index of indecision scores are also reported in the manual. Mean values for index of indecision and index of discrepancy are given for both high school seniors and college seniors. The college senior sample was composed of N = 158 individuals of both sexes and was drawn from Hiram College in Ohio and Huntingdon College in Alabama. Both colleges were coeducational, liberal-arts institutions with enrollments under 800 at the time the sample was drawn. College senior statistics are based upon pooled data. The high school senior sample totaled N = 165 individuals of both sexes, and was drawn from Shaker Heights high school in Ohio and Greenville high school in Alabama. Results from the two schools were pooled for the purposes of determining statistics for high school seniors. The test manual fails to note when the norms presented were obtained. However, much of the relevant data was gathered in 1949, and so it seems likely that norms were determined during the same period.

Availability:

The test appears to be out of print, but was originally available from Psychometric Affiliates, P.O. Box 3167, Munster, Indiana. The most recent catalog of Psychometric Affiliates does not list the test among those available.

Comments:

The Evaluation Aptitude Test is an interesting attempt to measure, in part, the influence of emotionally-loaded material on the ability to judge the validity of conclusions. The test is easily administered and scored, and potentially provides considerable information about an examinee's syllogistic reasoning ability. Sell is to be commended for including a variety of reliability and validity information in the test manual, which contributes greatly to an informed appraisal of the appropriateness of the test for a particular use.

However, the instrument suffers from a number of defects. Strictly speaking, the test does not appear to be an aptitude measure at all, as suggested by its title. Sell's statement that the instrument is "...meaningless for any person who has had training in deductive logic" (manual, p. 2) suggests that his test is more a measure of achievement than aptitude. Further, the fact that all the test items are syllogisms and that Sell several times refers to the test as a measure of deductive reasoning ability both suggest that the term "evaluation" appearing in the test title is narrowly defined by this test. Even "deductive reasoning" may be too broad a term, since the test does not include any conditional reasoning items.

The construction of the test also represents a potential source of problems. Part A and Part B appear opposite each other on the same page, and are arranged in such a way that logically equivalent syllogisms are also opposite each other. An examinee who recognized this pattern could complete Part A and transfer his responses directly across the page to the corresponding, logically equivalent syllogism in Part B. Such tactics would result in an index of discrepancy score of zero, suggesting that this particular examinee was unaffected by the emotional content of the Part B syllogisms, and could reason equally well about emotionally-neutral or emotionally-loaded material.

Probably the most serious problem of the test is the validity information presented in the manual. Part B appears to be easier than Part A, as evidenced by a higher mean score on Part B for college seniors, high school seniors, and 44 college subjects receiving the Evaluation Aptitude Test as a pretest prior to instruction in logic. If the emotional content of the Part B items significantly affects syllogistic reasoning ability, we should expect mean scores for Part B to be lower than mean scores for Part A. Yet Sell argues that the significant difference between Part A and Part B scores (favoring Part B) of the 44 college subjects "...can be attributed to nothing other than the nature of the syllogistic material of Part B and thus emotional bias -- since atmosphere effect, item difficulty and chance were held constant" (manual, p. 2). In the absence of a more complete explanation, this reviewer feels that the validity of Part B (and hence the test as a whole) has yet to be clearly established, since Sell's own results are contrary to what seem to be reasonable expectations.

Comments (continued):

And it also seems odd that a study such as the one Sell reports would be undertaken at all if, as he claims, the test is "...meaningless for any person who has had training in deductive logic" (manual, p. 2). If Sell is taken at his word, then scores on the posttest administration cannot be used to support his claim that the remaining discrepancy is due primarily to the continued operation of emotional bias, since such scores would be meaningless for this group of subjects who received one semester of logic instruction. It remains unclear what bearing (if any) this study has on establishing the validity of the test.

In addition, the intercorrelations reported for the five scores on the Evaluation Aptitude Test and five criteria of job success for the sample of 64 business agents yield no clear interpretation. It would be helpful to know, for example, why this particular group was studied, in order to judge the relevance of the findings to the validity of the test. Without such information, it is not clear how the reported intercorrelations should be interpreted and how they might contribute to the validity argument.

The norms presented in the test manual should probably be cautiously interpreted. The sample sizes seem too small for norming purposes, and the college senior sample is not likely to be a representative one. In addition, the norms may well be out-of-date.

In general, the attempt to measure the influence of emotionally-loaded material on syllogistic reasoning ability is noteworthy and worth refining. The most serious difficulty with the Evaluation Aptitude Test, however, is the lack of a clear validity argument. Until the validity of the test is established, the fairly high reliability estimates presented in the manual are of limited value in appraising the worth of the instrument. It is unfortunate that the validity of the Evaluation Aptitude Test remains doubtful; the instrument has clearly been constructed with considerable forethought and ingenuity, and potentially samples an important aspect of critical thinking ability.

Additional information:

Buros, O. K. Fifth mental measurements yearbook (1959). Reviews appear on pp. 785-787.

Ennis, R. H., & MacMillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Test citation:

Sell, D. E. Evaluation Aptitude Test. Munster, Indiana: Psychometric Affiliates, 1952.

INDUCTIVE REASONING TEST

Author(s): G. Bernard Baldwin
Year of publication: 1946
Most recent edition: 1946

Description:

The Inductive Reasoning Test consists of 40 scored and 4 sample items, and is designed to measure the ability "...to discover underlying principles in an arrangement of facts" according to the directions for test administration provided with the instrument. Examinees are presented with a number series in each test item, and must select the next two terms of the series from the five alternatives provided. In order to receive credit for an item, an examinee must correctly identify both succeeding terms. Test items are arranged in order of increasing difficulty. A time limit of 20 minutes is specified, but the author states that increasing the time limit to one hour increases the mean of large samples only one and one-half scores, which suggests that the test is essentially a power test. Scoring is by the 'rights only' method.

The test manual consists of a single page and contains information on reliability and validity estimates, development of the instrument, administration and scoring, and norms based on a sample of 3917 students in grades 9-12. The first, third, 16th, 50th, 84th, 97th, and 99th percentile points are identified. No specific recommendations are made concerning the grade levels for which the test might be appropriate.

Sample test item:

3 13 23 30 31 ______________ 41 32 33 42 43

[The initial series is given on the left. The alternatives from which the next two terms must be selected appear on the right.]

Reliability and validity information:

Odd-even reliability estimates were computed on the basis of a sample of 304 seniors from seven high schools in eastern Washington (state). The manual does not report these estimates, and states only that the reliability of the total test score is .93 for a single grade range. This figure may or may not represent the average reliability obtained from the sample of 304 seniors. No reliability estimates for specific grade levels are reported.

When the test was administered to 10 mathematics classes in Spokane, Washington at the beginning of the fall semester in 1940, the average correlation of total test score and instructor rankings of students' reasoning abilities was found to be .62. Correlations ranged from r = .52 to r = .85. The instructors were not informed of their students' scores on the test at the time the rankings were made, which was at the conclusion of

Reliability and validity information (continued):

the fall semester. For another sample of Washington high school seniors (N = 36), total test score correlated .69 with grade points in sciences and mathematics.

Test items were selected on the basis of ability to discriminate between an upper and a lower criterion group at each of two grade levels. The grade 9 group consisted of 160 beginning algebra students in a Washington high school. The grade 12 group was the same group on which reliability estimates were based. A sub-sample of students scoring below the first quartile (the lower criterion group) and a sub-sample of students scoring above the third quartile (the upper criterion group) were drawn for each grade level, and were used for the purposes of item analysis. By comparing the percentage of students in the upper criterion group who answered a given item correctly with the percentage of students in the lower criterion group who correctly answered the same item, an index of discrimination was obtained for a given item and a particular grade level. Large differences reflect greater discrimination ability of individual items. Baldwin states that the initial 32 items of the test provide good discrimination for grade 9 students, and the last 32 items discriminate highly among high school seniors.

Availability:

The test is now out of print.

Comments:

The Inductive Reasoning Test is brief, easily-administered and easily-scored. However, it suffers from a number of weaknesses. Although Baldwin states that the instrument is designed "...to measure [inductive reasoning ability] as independently of language skills as possible," performance on the test would seem to be heavily influenced by numerical ability. In the sample test item reproduced here, for example, a student with limited mathematical background might not recognize that the terms in the number series might easily be interpreted as numerals in base 4 — and the answer then becomes predetermined, since only two of the alternatives represent legitimate base 4 numerals (32 and 33). Further, Baldwin's validity estimates are all based on samples of mathematics students; the obtained correlations might be accounted for in terms of mathematical ability rather than inductive reasoning ability.

The test also suffers from a problem common to tests of inductive reasoning. More than one pattern can often be abstracted from a given series. In the sample test item, adding together the digits of each term produces the following series: 3, 4, 5, 3, 4, ______________. This series might be completed by the alternatives 41 and 42, 41 and 33, or 32 and 33,

Comments (continued):

each of which completes the initial series in exactly the same way, although the keyed answer is 32 and 33. By continuing the addition of digits, the following series is produced, regardless of which of the three above pairs is selected: 3, 4, 5, 3, 4, 5, 6. A student who had discovered the pattern sketched here and who happened to select 41 and 42 or 41 and 33 as the correct response would be penalized. It should be noted, however, that 14 test items include the next three terms of a particular series among the alternatives. Requiring examinees to identify the successive three terms of a series might alleviate the problem of multiple patterns to some extent.

Finally, it seems questionable that the test could provide a measure of inductive reasoning ability in any general sense. The test items are mathematically-oriented and are isolated from any context. The test format and short time limit both encourage examinees to abstract only one pattern from the given number series, rather than generating additional patterns from which the series may be derived with equal plausibility. The validity argument presented in the manual is unconvincing; Baldwin's confounding of inductive reasoning ability and numerical facility should serve as a signal that test results probably do not admit of any clear interpretation.

Additional information:

Buros, O. K. The third mental measurements yearbook (1949). Review appears on pp. 314-315.

Ennis, R. H., & MacMillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Test citation:

Baldwin, G. B. Inductive Reasoning Test. Minneapolis, Minnesota: Educational Test Bureau, 1946.

INTERPRETATION OF DATA TEST
(Progressive Association Test 2.52)

Author(s): Evaluation Staff, Eight-Year Study
Year of publication: 1940
Most recent edition: 1940

Description:

The Interpretation of Data Test was developed as an evaluation instrument in the Eight-Year Study, and also served as the model for the construction of the Illinois Test of Ability to Judge Interpretations of Data (also reviewed here). The Interpretation of Data Test is intended to assess examinees' proficiency in two general areas: (1) ability to perceive relationships in data; and (2) ability to recognize the limitations of data (Smith and Tyler, 1942).

The test presents examinees with information contained in various charts, tables, graphs and prose passages. Each presentation of data is followed by a series of 15 statements. For each statement, examinees must judge whether the data provided are alone:

(1) sufficient to make the statement true
(2) sufficient to indicate that the statement is probably true
(3) insufficient to indicate whether there is any degree of truth or falsity in the statement
(4) sufficient to indicate that the statement is probably false
(5) sufficient to make the statement false

Examinees are required to evaluate 150 statements in the course of the test. Statements are classified by the type of ability or behavior required to evaluate the data addressed by the statement. These classifications are as follows: (1) comparison of data points; (2) comparison of trends in the data; (3) perceiving major trends — generalization; (4) interpolation; (5) extrapolation; (6) sampling; (7) recognition of statement of purpose; (8) recognition of analogy being drawn; (9) identification of cause; (10) identification of effect; and (11) recognition of value statement. At least one of each type of statement is contained within each group of 15 statements which follow a particular presentation of data.

A number of scores are calculated for this test. The general accuracy score is computed as the total number of correct responses, using the 'rights only' method. This score is further subdivided into three scores: (1) accuracy on statements keyed as 'probably true' and 'probably false'; (2) accuracy on statements keyed as providing insufficient data; and (3) accuracy on statements keyed as 'true' or 'false'. The beyond data score is computed as the total of 'true' responses to items keyed as 'probably true', 'false' responses to items keyed as 'probably false', and 'probably true' or 'probably false' responses to items keyed as 'insufficient data'. This score is intended to indicate the willingness of an examinee to grant greater justification to a statement than is warranted by the data.

Description (continued):

The caution score is computed as the total of 'probably true' responses to items keyed as 'true', 'probably false' responses to items keyed as 'false' and 'insufficient data' responses to items keyed as 'probably true' or 'probably false'. This score is intended to indicate the extent to which examinees are overcautious in evaluating statements.

Omitted items are not scored as incorrect, but are used as a basis for determining the number of opportunities an examinee had to respond correctly to items of that same type. The crude error score is determined by adding the number of 'false' and 'probably false' responses to items keyed as 'true' to the number of 'true' and 'probably true' responses to items keyed as 'false'. This type of error is assumed to be due to carelessness, misunderstanding, or inability to read the data correctly.

No time limit is specified for completion of the test. The instrument is intended to be used with senior high school students. A score interpretation and evaluation booklet accompanies the test, but no reliability or validity information is provided. However, Smith and Tyler (1942) present some reliability and validity information for this measure.

Note:

Problems from Test 2.52 and Test 2.51 (an earlier form of the Interpretation of Data Test 2.52) were combined in constructing the Interpretation of Data Test -- Upper Level, published by the Cooperative Test Division of the Educational Testing Service in 1950. The two forms of the ETS instrument follow the same format as Test 2.52, except that each form consists of 75 rather than 150 items. Essentially, Test 2.52 and Test 2.51 were developed as two half-tests.

Another Progressive Association Test (Test 2.71) served as the basis for the development of the Interpretation of Data Test -- Lower Level, also published by the Cooperative Test Division of ETS in 1950. Again, the original Test 2.71 was divided into two halves, each with 50 items. The format of the test is similar to the Upper Level version, with the exceptions that only 10 statements follow each presentation of data, and that examinees must judge whether: (1) enough information is given to make the statement true; (2) enough information is given to make the statement false; or (3) not enough information is given to warrant a decision.

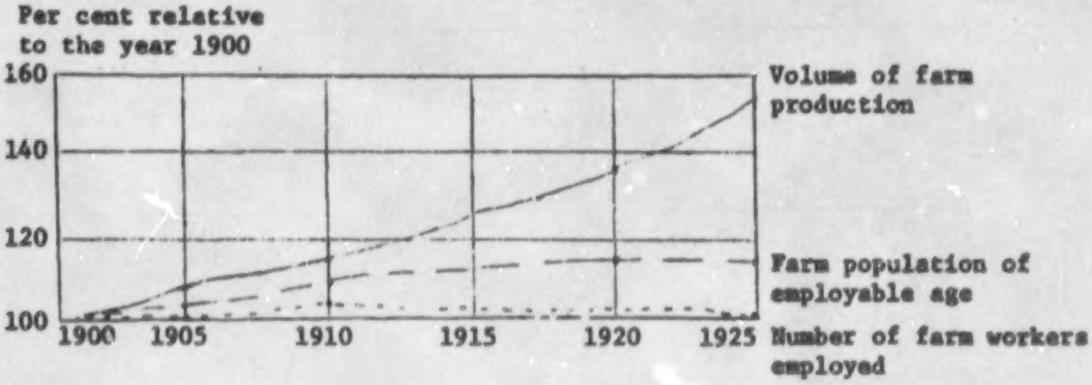
For the purposes of this review, only the original Progressive Education Association Test 2.52 will be discussed in any detail. The same scores are computed for all the measures listed above (although the method of computation varies somewhat as a function of the complexity of the possible responses). Because the later ETS instruments were modeled so closely after the original PEA versions, the following discussions should be sufficient to acquaint

Note (continued):

the reader with the general approach taken in developing and interpreting these measures. Interested readers are referred to the specific instruments for comparisons and evaluations which may be of concern.

Sample test item:

Problem I. This chart shows production, population and employment on farms in the United States for each fifth year between 1900 and 1925.



Statements

1. The ratio of agricultural production to the number of farm workers increased every five years between 1900 and 1925.

2. The increase in agricultural production between 1910 and 1925 was due to more widespread use of farm machinery.

....

6. During the entire period between 1900 and 1925 there was an excess of farm population of employable age over the number of people needed to operate farms.

....

12. The average amount of farm production was higher in the period 1925 to 1930 than in the period 1920 to 1925.

13. Between 1900 and 1925, there was an increase in the farm population of employable age in the Middle West, the largest farming area in the United States.

Statements (continued)

14. Farm population of employable age was lower in 1930 than in 1900.

15. The production of wheat, the largest agricultural crop in the United States, was as great in 1915 as in 1925.

Reliability and validity information:

Smith and Tyler (1942) present considerable reliability and validity information for this test. In discussing the Interpretation of Data Tests 2.52 and 2.71, they consider two aspects of validity: (1) the validity of the instruments as measures of the ability to judge interpretations of data formulated by others; and (2) the validity of the instruments as measures of the ability to create original interpretations of data.

Smith and Tyler argue that the validity of Test 2.52 as a measure of the ability to judge interpretations of data is supported by the test construction procedures, which yielded items which provide examinees with the opportunity to display behaviors which indicate proficiency in perceiving relationships in the data and in recognizing limitations of the data. Further, they claim that the 5-choice response scale for each item requires an examinee to respond in such a way that the extent of his recognition of both relationships and limitations in the data may be inferred from his response. Smith and Tyler also note that the validity of the test is supported by the expert judgment of members of the Evaluation Committee in determining the criteria by which items were selected. On the basis of such judgments and an analysis of the types of data normally found in curricular materials, the items comprising the test are claimed to represent an adequate sample of the types of data interpretation problems likely to be encountered by senior high school students.

The second validity consideration was addressed in two studies of grade 9-12 students. One group (N = 119) was drawn from a large public high school and the other group (N = 99) was drawn from a smaller, private high school. Both groups were administered an essay form of Test 2.52 consisting of 42 questions. Seven questions were formulated for each of six exercises in Test 2.52, and each question corresponded in subject matter and type of relationship to a statement used in Test 2.52 (i.e., the essay questions required abilities or behaviors corresponding to those required by particular statements on Test 2.52 on which the essay questions were based). Only 39 questions were scored as 3 questions were later found to be ambiguous. Essay responses were scored to yield 5 scores corresponding to the scores computed for Test 2.52: (1) accuracy; (2) beyond data; (3) crude error; (4) caution; and (5) unable to perceive relationship in data. Essays were scored independently by three raters following explicit guidelines. Interrater reliabilities were reported as ranging from r = .94 to r = .96.

Reliability and validity information (continued):

Approximately one week after administering the essay form of Test 2.52, the regular form was administered to both groups, and corresponding part scores for each test were correlated. Separate correlations were computed for each group. Correlations ranged from r = .08 (crude error; large public school) to r = .74 (general accuracy; large public school), uncorrected for attenuation. Corrected coefficients are also presented for each group.

Smith and Tyler also indicate that 19 samples of grade 9-12 students in a total of 7 schools were used to estimate the reliability of Test 2.52. However, they present data only for the groups which participated in the validity study cited above. For these groups, Kuder-Richardson reliability estimates ranged from r = .75 (crude error; small private school) to r = .95 (general accuracy; large public school).

Means and standard deviations of per cent scores for part scores for the above two groups are also presented.

Availability:

The Interpretation of Data Test (Test 2.52) is out of print. The ETS versions of this instrument are also out of print.

Comments:

The Interpretation of Data Test is an ambitious attempt to assess examinees' abilities to evaluate data. The instrument potentially provides considerable information about specific strengths and weaknesses of examinees in appraising data. However, the measure suffers from a number of problems.

Only a minimum of evidence is offered to support the validity of the test. The expert judgment which served as the basis for item selection supports only the content validity of the instrument, and not other types of validity which are also important. The reported correlations of Test 2.52 with an essay form of the same test cannot be used as evidence of the validity of either test, since neither instrument was satisfactorily validated before being used in the validity studies. In addition, Smith and Tyler (1942) make it clear that the essay form of the test was used as the criterion, even though the validity of this form had been investigated even less than the validity of Test 2.52.

Further, the high reliability estimates should be interpreted cautiously. A test with 150 items should have at least respectable reliability estimates,

Comments (continued):

simply due to its length. The high reliability estimates reported for this test may be due primarily to the length of the test, rather than accuracy of measurement.

The most serious problems with the test, however, center on the items themselves. For example, in the sample test problem reproduced here, statement 12 is keyed as 'probably true', presumably on the assumption that the trend of generally increasing volume of farm production is likely to continue. However, the graph presents no information whatsoever about the years 1925-1930, and in the judgment of this reviewer, the data provided are not alone "sufficient to indicate that the statement is probably true," which is the criterion to be met in judging a statement to be 'probably true', according to the test directions (see p. 1). In addition, statement 13 is keyed as 'probably true', presumably on the assumption that what is true of the whole is likely to be true of a large portion of the whole. However, no information is presented which might justify the assumption that the Middle West is representative of the United States farming community; without this assumption, the data presented do not seem sufficient to indicate that statement 13 is probably true. Statements 14 and 15 suffer from similar problems. The keyed answers cannot be supported without additional assumptions, and this fact alone should be sufficient evidence that the data presented are not alone sufficient to determine the truth value of the statements. And the data provide only limited (at best) justification for making the required assumptions, and so such assumptions must be considered questionable. For these reasons, this reviewer feels that the best answers to the items mentioned above are all 'insufficient evidence'.

Examinees who reason similarly will achieve lower general accuracy scores and higher caution scores, which may not adequately reflect their abilities to evaluate data. It is also unfortunate that these same difficulties have been carried over on the ETS versions of this test (advanced level). Both the PEA and ETS versions of this instrument could probably benefit from a logical analysis of the test items and keyed answers.

In general, the Interpretation of Data Test is a potentially useful instrument capable of isolating specific difficulties of examinees in evaluating data. It suffers, however, from the lack of a strong validity argument and logical problems inherent in the test items themselves. The approach to assessing the aspects of critical thinking ability sampled by the measure is noteworthy and worth refining. In its present form, however, the Interpretation of Data Test should probably be considered promising, but incomplete.

Additional information (for both PEA and ETS versions):

Amstutz, W. S. A study of a test for measuring skill in the interpretation of data. Journal of Experimental Education, 1940, 8, 251-255.

Buros, O. K. The third mental measurements yearbook (1949). Review appears on pp. 561-562.

Buros, O. K. The fourth mental measurements yearbook (1955). Reviews appear on pp. 612-615.

Smith, E. R., & Tyler, R. W. Appraising and recording student progress. Adventure in American Education, Volume III. New York: Harper and Brothers, 1942.

Test citation:

Evaluation Staff, Eight-Year Study. Interpretation of Data Test. Chicago: Progressive Education Association, 1940.

Author(s): Barbara Moor Sanner, et al.
Year of publication: unpublished
Most recent edition: 1972

Description:

The Illinois Conditional Reasoning Test was developed as part of Sanner's dissertation research, and incorporates a number of ideas taken from the Smith-Sturgeon Conditional Reasoning Test (also reviewed here). Sanner's instrument is designed to assess mastery of four conditional reasoning principles: (1) particular contraposition (valid); (2) particular transitivity (valid); (3) particular conversion (affirming the consequent -- invalid); and (4) particular inversion (denying the antecedent — invalid).[1] The test consists of 24 items, divided equally among the four principles. In addition, the type of premise is varied within the six items testing for each principle. Three items are presented in factual form (i.e., the premises would likely be taken as true by the examinee) and three items are presented in suppositional form (i.e., the premises would likely be taken either as false or of uncertain truth-value). The content of the test items is classified as familiar, that is, the examinee could reasonably be expected to have knowledge or beliefs about the material given in the item. Test items were also designed so that examinees would be uncommitted to the truth or falsity of the conclusions reached in the arguments.

The format of the test departs considerably from the standard paper-and-pencil type of test. The test is administered individually, and requires one examiner and one scorer. The conditional reasoning principles are illustrated through the use of simple chemical reactions demonstrated to the examinee as the test is administered. For each section of the test, examinees must first master the principle of detachment for administration to continue. For example, in order to complete the first six items, the examinee must demonstrate understanding of the conditional statement, "If vinegar is added to [baking] soda, then the soda bubbles." This statement serves as the major premise for items in this section. Mastery of the principle of detachment is established when the examinee can consistently predict the correct result of adding vinegar to soda. The principle is demonstrated to the examinee and reinforced visually with a picture board with interchangeable drawings which represent the various manipulations taking place. Examinees must master one major premise in sections 1, 4 and 6; in sections 2, 3 and 5, two major premises must be mastered, since those sections introduce the transitivity items, which require three premises. If, after several attempts, an examinee fails to understand the appropriate conditional statement(s) (as evidenced by failure to master the principle of detachment for that section), the test section should be skipped. For factual items in each section, the minor premise is demonstrated to the examinee, using the actual materials involved. For suppositional items, the examinee is asked to pretend that the minor premise is true, but does not receive a demonstration.

---

[1] Whenever possible, the logical principles discussed in this review are classified by the system suggested by Ennis (1976).

Description (continued):

Scoring is somewhat complicated, and requires the use of a special answer sheet. Both part and total scores can be computed. Examinees must provide both the correct response to the test item and an adequate justification in order to receive credit. Scoring is by the 'rights only' method. From the examinee's viewpoint, test directions are fairly simple, and require only that he or she observe what is occurring and respond to oral questions. From the examiner's viewpoint, however, test administration is considerably more complicated, as the examiner must read the script supplied with the test and manipulate the appropriate experimental materials, while another person scores examinee responses (which sometimes will require additional probing and thoughtful judgment in order to be evaluated). Sanner recommends against one person serving as both examiner and scorer.

The author states that the test is appropriate for use in grades 1-10, but also cautions that the test administrator must be the ultimate judge of the appropriateness of the test (the test might be inappropriate for students who have a substantial knowledge of chemistry, for example). Sanner specifically notes that the test is untimed; test administration typically requires 30-40 minutes for children between 8 and 11 years, according to Sanner's dissertation.

Although no manual accompanies the test, Sanner's dissertation includes information on the development of the test, instructions for administration and scoring (including judging justifications offered for responses), and reliability and validity information. Only 50 students from a single school in California were used in Sanner's research; hence, no norms are reported. According to Sanner's description, the group used for the study was composed of 50 grade 5 students from a single school in a predominantly middle-class neighborhood, Mexican-American students comprised 22% of the sample, and Blacks, 2%. Ages ranged from 10 years 6 months to 12 years 10 months. Boys and girls were approximately equally represented in the sample. Students came from a wide range of socio-economic backgrounds, and showed considerable variation in I.Q. scores. Mean verbal I.Q. was 98.56 with a standard deviation of 13.67; mean non-verbal I.Q. was 106.95 with a standard deviation of 14.64. These results are based upon the 41 students for whom I.Q. data was available.

Sample test item:

[Capitalized words are those spoken by the examiner. Quoted words are responses (or possible responses) given by the examinee.]

1. DOES VINEGAR MAKE SODA BUBBLE?

   [checks mastery of major premise for this section: If vinegar is added to soda, then the soda bubbles.]

2. I'M GOING TO ASK YOU TO PRETEND SOMETHING. THEN I WANT YOU TO TELL ME WHAT WE'RE PRETENDING. PRETEND THERE'S A WHITE POWDER IN HERE [a beaker]; WE ADD VINEGAR AND IT DOESN'T BUBBLE. WHAT ARE WE PRETENDING?

   [checks mastery of minor premise: Vinegar is added to a white powder and no bubbles result.]

   "We're pretending that we added vinegar and the white powder didn't bubble."

3. WE'RE PRETENDING WE ADDED VINEGAR AND THE POWDER DIDN'T BUBBLE. DO YOU THINK THE POWDER WOULD BE SODA, WOULDN'T BE SODA, OR IS THERE NO WAY TO TELL?

   a. "Would be soda"

   b. "Wouldn't be soda"

   c. "There's no way to tell"

   [checks mastery of particular contraposition principle for suppositional item; note 'pretend' in lines 2-3]

4. TELL ME HOW YOU DECIDED THAT...

   [requests justification as additional check on mastery of principle]

Reliability and validity information:

Kuder-Richardson 20 reliability estimates for both part and total test score were computed on the basis of 49 students (one of the original 50 students was eliminated due to equipment failure). Reliability for total test score was estimated to be .77. Reliability estimates for part scores ranged from r = .51 (suppositional items) to r = .94 (invalid principle items). Test-retest reliability estimates for both part and total test scores were computed on the basis of a random sample of 10 of the 49 experimental subjects. The interval between test administrations varied between 5 and 12 days. Test-retest reliability estimates for part scores ranged from r = .50 (factual items) to r = .90 (conversion items). All part scores for the four conditional logic principles showed a significant test-retest correlation at the .05 level.

Reliability and validity information (continued):

Sanner also computed a rescoring correlation using randomly selected tapes of the performances of six individuals in the study. The scoring-rescoring correlation was quite high; the minimum correlation obtained was $r = .97$ (suppositional items).

Sanner argues that the content validity of the Illinois Conditional Reasoning Test is supported by an examination of the script and scoring sheets of the instrument, and the fact that the correct response to each test item is, in fact, implied by the conditional reasoning principle relevant to that item. Construct validity of the test was weakened by a failure to find significant hypothesized correlations between principle scores and I.Q. However, a few significant correlations were found: (1) $r = .35$ for total test score and non-verbal I.Q.; (2) $r = .38$ for suppositional items and non-verbal I.Q.; and (3) $r = .33$ for suppositional items and verbal I.Q. These correlations were significant at the .05 level. All other correlations were non-significant, although they were in the hypothesized direction (positive).

Availability:

The test is unpublished, but copies of Sanner's dissertation which includes instructions for administration and scoring as well as a copy of the test itself, are available from University Microfilms, order number 74-14,609.

Comments:

The Illinois Conditional Reasoning Test is noteworthy for its imaginative approach to assessing mastery of four conditional reasoning principles. The language used in the test seems appropriate for the grade levels for which the test is designed, with the qualification that older examinees might not need to repeat what is being pretended in suppositional items (cf. sample test item). The script provided with the test is generally clear in its directions to the test administrator, and should help to standardize test administration by insuring that each examinee is asked identical questions and presented with the same demonstrations. The use of the picture board to reinforce the major premise(s) of each test section represents an attempt to eliminate poor performances due to misunderstandings or forgetting of the basic information given in the problem. This practice might also prove unnecessary for older students.

For the purposes of the study in which this instrument was used, Sanner operationally defined probable mastery of a principle as at least five correct responses (i.e., correct response with adequate justification) out of the six items for each principle. Three or fewer correct responses for a single principle is given as the criterion for probable non-mastery of the principle.

Comments (continued):

No mastery statement is made for a score of four correct. Both part and total scores can be computed, although this reviewer feels that the test-retest reliability estimates for part and total scores reported in Sanner's dissertation must be viewed cautiously, since they are based upon a sample of only 10 students. The Kuder-Richardson reliability estimates reported by Sanner are estimates of internal consistency rather than consistency of performance over time. For this reason, it is difficult to judge the likelihood of obtaining relatively consistent results when computing part and total test scores from this instrument. The construct validity argument might profit from additional investigations of the correlations between performance on the Illinois Conditional Reasoning Test and performance on other measures. However, despite the weakness of the construct validity argument, the test appears to have a reasonable degree of both face validity and content validity.

One drawback to this instrument is the complicated administration and scoring procedures. Administration requires a variety of special materials (beakers, chemicals, etc.) and is individual. Scoring is complicated, since responses are scored in three parts: conclusion, justification and final score. Conclusions may be sub-classified into four types and justifications sub-classified into ten types, and both may be coded for more efficient scoring. The increased time and effort required for administration and scoring should be evaluated against the accuracy of assessment of mastery potentially provided by the test, however. By requiring examinees to provide both a correct answer and an adequate justification in order to receive credit for an item, correct responses due to guessing should be considerably reduced (and likewise for correct responses given for wrong reasons). An empirical investigation of this hypothesis might prove informative; if examinees tended to obtain the same score on the Illinois Conditional Reasoning Test whether or not they were required to justify their answers, then the test might profitably be used in a video-tape format, which would allow more control over visual presentation and mass administration as well.

Additional information:

Sanner, B. M. A study of the ability of fifth graders to handle conditional logic (Doctoral dissertation, University of Illinois, 1974). Dissertation Abstracts International, 1974, 35, 111. (University Microfilms No. 74-14, 609)

Test citation:

Sanner, B. M., Ennis, R. H., Martin, M., Smith, E. L., & Sturgeon, J. Illinois Conditional Reasoning Test. Unpublished test. Urbana, Illinois: University of Illinois, 1974.

Author(s): J. Thomas Hastings, David Krathwohl, and Dora Damrin
Year of publication: unpublished
Most recent edition: 1954

Description:

The Illinois Test of Ability to Judge Interpretations of Data is intended to assess the ability to judge "...to what extent given interpretations of data are justified by the data," according to the statement on the test booklet cover. The instrument consists of two parts, with five problems in each part. In each part, examinees are presented with information organized in charts, tables or graphs and must evaluate the truth of statements which follow. Ninety statements must be evaluated in each part of the test. Examinees are asked to assume that the information presented is accurate and to evaluate the truth of the statements only on the basis of the information provided. The possible responses are as follows: (1) the statement is true; (2) the statement is probably true; (3) the information presented is insufficient to determine the truth-value of the statement; (4) the statement is probably false; and (5) the statement is false. In Part I, the statements to be evaluated are presented opposite the table, chart or graph in a list. In Part II, the statements appear in context (usually a speech or editorial) and are numbered for reference in the text. The intent of Part II is to measure the extent to which reasoning abilities are influenced when data are presented in an emotional context.

Several scores are computed for this test. The "data reading" score is designed to reflect the ability of an examinee to comprehend the information presented in the table, chart or graph. Data reading items are those which restate information presented graphically; all of the data required to judge the truth or falsity of the statement is given in the chart, table or graph. Twenty-one items are classified as data reading items in each part of the test. A low score indicates difficulty in reading data correctly, and the authors note that other scores will be affected, since errors in reading, even when accompanied by careful thinking, will result in poorer performance on other test items. The "overcaution" score is intended to reflect the extent to which an examinee is unwilling to accept statements made tenable (but not established) by the data. The score is computed by summing the number of "insufficient data" responses to items keyed as "probably true" or "probably false". Higher scores indicate greater caution. The "beyond data" score is an indication of an individual's willingness to accept a statement not justified by the data; such statements are often value judgments. Higher scores indicate a tendency to attribute greater justification to statements than is warranted by the data. The total test score is intended to reflect the examinee's general competence in judging interpretations of data.

Total test scores and data reading scores are classified as "accuracy scores" and are computed as the sum of correct responses. Overcaution and beyond data scores are classified as "error scores" and require somewhat different calculations, which are described in the manual. Overcaution,

Description (continued):

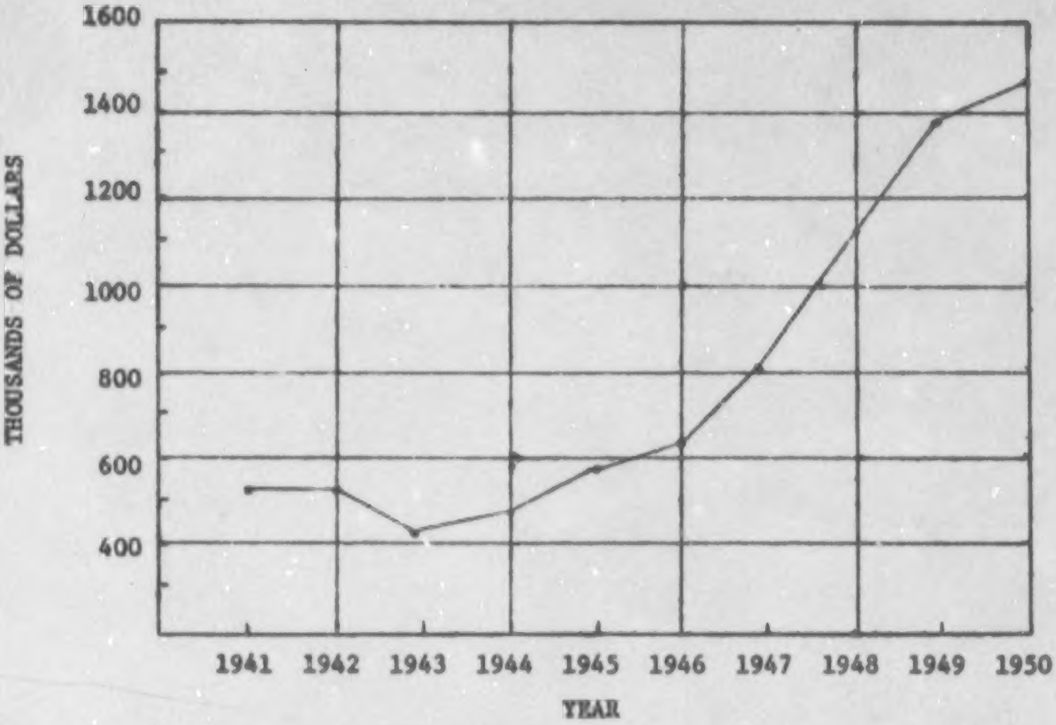
beyond data, and total test scores are calculated for Part I and Part II of the test. Changes in these scores between test parts are alleged to provide a measure of the influence of emotional contexts on reasoning abilities. The data reading score is computed only for Part I of the test, presumably on the assumption that if an examinee cannot read and understand the data presented in isolation (i.e., out of any context) then he will not be able to read and understand data in an emotional context.

Test administration time is lengthy. According to the manual of directions, approximately 38 minutes are required to familiarize students with the directions for taking the test and working a sample problem for Part I. Part I of the test has a specified time limit of 50 minutes; in addition, examinees are limited in the amount of time spent on any one of the five problems in Part I. Hence, the test administrator must monitor the elapsed time closely. Part II of the test requires approximately 20 minutes for the directions to be explained and examinees to work through the sample problem. A time limit of 64 minutes is given for the five problems in Part II; the individual problems are also subject to a time limit. Examinees are instructed to make their best judgment on items for which they are uncertain in both parts of the test. According to the manual, the test is designed to be administered in two sessions on the same day, but may be given in two sessions on consecutive days. A 15-20 minute break is recommended if the test is administered in two sessions on the same day.

The test is recommended for use in grades 11-12, although it has been used with college students on an experimental basis. The manual which accompanies the test presents norms for two consecutive years (1953 and 1954) which are essentially identical. The 1953 norms are based upon a sample of 400 students; no sample size is reported for the 1954 norms. A discussion of test score interpretation is also provided in the manual, and includes hypothetical examples of different score configurations and their corresponding interpretations. The manual contains no validity or reliability information. A separate manual for test administration gives explicit directions for administering the test.

Sample test item (Part I):

AMOUNT SPENT PER YEAR BY THE U.S. GOVERNMENT
ON A PROGRAM OF PLANTING FOREST TREES



5. More trees were planted under this program in the last six months of 1945 than during the similar period in 1946.

Reliability and validity information:

The test manual presents no reliability or validity information. In addition, reliability estimates for earlier forms of the instrument were computed for different part scores than those provided by the latest test form. These previous reliability estimates are therefore not directly comparable to any part score reliability estimates which might be computed for this form of the test.

Availability:

The test was never published. The instrument received considerable attention in the early 1950's, and was then abandoned. No further development is anticipated by the authors at this time, although the Illinois Critical Thinking Project is considering adopting the test.

Comments:

The Illinois Test of Ability to Judge Interpretations of Data is an ambitious attempt to assess the skills of examinees in evaluating conclusions drawn from data. The idea of measuring the influence of context on the ability to evaluate conclusions drawn from data is noteworthy and worth refining. Likewise, isolating specific difficulties in reasoning about conclusions drawn from data, such as the tendency to be overcautious or the tendency to go beyond the data, would seem to be a valuable function of the test.

The major difficulty with this instrument, however, is the assumptions about the data required in order to justify the keyed answers. In the opinion of this reviewer, students who are unwilling to make questionable assumptions about the data may be penalized by increased overcaution scores, and lower total test scores. Consider the following example (which refers to the sample test item given earlier):

The keyed answer to the sample test item is given as "probably false". The manual presents this rationale in defense of the keyed answer:

> The total amount spent in 1945 was not greater than the total amount spent in 1946; and since we would expect the last six months to follow the general yearly pattern, we can say that the item is probably false (pp. 3-4; emphasis in original).

Further, this rationale is presented to the examinees in the course of working the sample problem for Part I of the test. In effect, examinees are encouraged to make the sort of assumptions illustrated in the rationale for the keyed answer.

However, this reviewer feels that such assumptions are often unwarranted and need not reflect any critical thinking by the examinee. For this sample problem, the same graphic information could be generated by little tree-planting activity in the first six months of 1945, and much tree-planting activity during the last six months of 1945, with peak tree-planting activity occurring in the first six months of 1946, followed by a decline to levels below those of the latter part of 1945 during the last six months of 1946. Under this interpretation, the total number of trees planted in 1946 could still be greater than in 1945; but more trees would have been planted in the last six months of 1945 than during the last six months of 1946.

Further, it should be recognized that the graph presents no information whatsoever on the number of trees planted in various years. Rather, government expenditures on a tree-planting program are represented. Comparisons of the number of trees planted in given years can be made only on the assumption that greater expenditures reflect greater numbers of trees planted. But this assumption is also questionable. It seems likely that increased costs of planting might result in greater expenditure of funds for identical numbers of trees planted in consecutive years.

Comments (continued):

For the above reasons, this reviewer believes that the best answer to the sample item is "insufficient information". A logical analysis of these sorts of questions might eliminate such difficulties, and ultimately improve the validity of the instrument. Until such an analysis is completed, however, the keyed answers should remain suspect.

Test citation:

Hastings, J. T., Krathwohl, D., & Damrin, D. Illinois Test of Ability to Judge Interpretations of Data. Unpublished test. Urbana, Illinois: University of Illinois, 1954.

LOGICAL REASONING

Author(s): Alfred F. Hertzka and J. P. Guilford
Year of publication: 1955
Most recent edition: 1955

Description:

The Logical Reasoning test was developed on the basis of factor-analytic studies conducted by J. P. Guilford and other researchers. The currently available form of the test is described in the Sheridan Psychological Services 1976-1978 Catalog as measuring "...the ability to judge the logical soundness of meaningful conclusions; also, the trait that is popularly known as critical thinking ability" (p. 12). The manual refers to the trait measured by the test as the factor of "logical evaluation" (p. 3). The instrument consists of 40 items, and is divided into two parts each containing 20 items. Four practice items are provided at the beginning of the test. Each test item presents examinees with two premises and four possible conclusions. Examinees must choose from the four alternatives provided, the single conclusion which follows from both premises. The test authors state that the items cover "...the 15 valid syllogistic forms" (manual, p. 1). They also state that the syllogistic item format was selected because tests composed of items in this format have consistently shown the highest factor loadings on the ability purportedly measured by the test. Test items are intended to be "generally true and factual in the realistic sense" (manual, p. 1), and their content includes references to the fields of business, the sciences, sports and art. The authors state that the two parts of the test are parallel and may be administered as individual instruments (reliability will be lowered, however). Hertzka and Guilford also suggest that the items are scaled in difficulty for each test part, although they do not state this explicitly.

No recommendations are made concerning the age groups for which the test is appropriate. However, the manual includes reliability data and norms for both high school and college students. Validity information is reported for college students and students in Air Force Officer Candidate School.

A time limit of 10 minutes is specified for each part of the test. An additional 5 minutes should be allowed for examinees to read the test directions and to work the sample problems. The manual states that the time limit for each part of the test "should be rigidly adhered to" (p. 3). However, the manual also cautions test administrators to "...allow for at least 30 minutes total time for the complete test and 20 minutes for the half test" when testing small groups (under 30 examinees) (p. 4). Examinees are not permitted to return to Part I after beginning work on Part II of the test.

Scoring is done by a special correction-for-guessing formula (rights plus 1/4 of omitted items), which is designed to avoid the problem of negative scores generated by conventional correction-for-guessing formulas. According to the authors, their method of scoring yields scores which are "perfectly correlated with the score[s] that would be found by the more conventional correction-for-guessing formula" (manual, p. 4).

Description (continued):

Test directions are generally clear and straightforward. Examinees are encouraged to answer all items, although they are cautioned against wild guessing and informed of the penalty for guessing. The instructions explicitly state that every item has only a single correct answer, and the correct response is one "...derived from both statements, and from those statements only. A correct conclusion is not just a repetition of the contents of one of the statements. A correct conclusion is not based on information other than that supplied by the given statements" (test, p. 2; emphasis in original).

Sample test item:

All automobiles are cars.

All convertibles are automobiles.

Therefore:

A. Some automobiles are not convertibles.

B. Some cars are convertibles.

C. All cars are automobiles.

D. All convertibles are cars.

Reliability and validity information:

Reliability estimates were based upon stratified random samples of high school and college students (N = 100 for each). The manual provides reliability estimates for Part I, Part II, and the complete test. The method by which reliability estimates for Part I and Part II were obtained is not reported; the Spearman-Brown formula was used to compute an estimate of the reliability for the complete test, based upon the reliability estimates obtained for each part. Reliability estimates for the high school sample were as follows: r = .83 (Part I, Part II); r = .91 (total test). Reliability estimates for the college sample were reported as r = .80 (Part I, Part II); r = .89 (total test).

Norms for Part I and total test scores are presented for the larger groups from which the samples were drawn for the purposes of reliability estimation. The high school norm sample (N = 402) consisted of students in grades 10-12, of which about 54% were male. Ages ranged from 14-20 years. The college norm sample (N = 509) was composed of 38% freshmen, 34% sophomores, 16% juniors and 12% seniors. Males comprised 70.5% of this sample. Median age for the group was 21.6 years. Students in both samples were drawn from schools in San Diego, California. The manual does not

Reliability and validity information (continued):

state when the norms were obtained. Means, standard deviations and standard errors are also reported for each group on part and total scores.

The authors present limited evidence for the validity of the test. Factor analysis of the present form of this instrument revealed a loading of .50 in the factor identified as logical evaluation by Hertzka and Guilford. The test manual does not identify the research which produced this result. A similar form of this test was found to correlate r = .32 with GPA of 54 engineering students, and to yield a median correlation of r = .37 with scores on a mathematics proficiency test for a sample of 70 engineering students. At a different institution, scores on a similar form of the test correlated r = .14 and r = .04 with GPA in calculus, for two samples of N = 35 and N = 41 mathematics and physics students. At this same institution, score on a test similar to the Logical Reasoning test was found to correlate r = .42 with grades in advanced calculus (N = 12). All of these results were obtained by John R. Hills in his dissertation research (referenced below). In addition, the manual also reports correlations between scores on Thurstone's False Premises, Inferences, Syllogisms and Word Matrix tests and four criteria of achievement for a group of N = 114 students in Air Force Officer Candidate School. The authors state that Thurstone's tests had high loadings on the factor of logical evaluation.

Availability:

The test is currently in print and is available from Sheridan Psychological Services, Incorporated, P.O. Box 6101, Orange, California, 92667 to those individuals who meet the qualifications outlined in the Sheridan Psychological Services Catalog. The catalog should be consulted for complete information.

Comments:

The Logical Reasoning test appears to be a straightforward attempt to assess a single aspect of critical thinking ability. Although the test is variously described as containing heavy loadings on the "logical evaluation" factor, and as a "measure of the factor evaluation of semantic implications" (Sheridan Psychological Services Catalog, 1976-1978, p. 12; emphasis in original), these descriptions are somewhat misleading. All test items require examinees to apply principles of class reasoning in identifying correct conclusions; no conditional reasoning is required. As a test of class reasoning ability, the test may have some value; there is, however, no basis to think that the test measures the trait "popularly known as critical thinking ability".

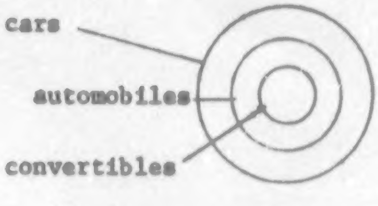
Comments (continued):

Aside from this difficulty, the test suffers from several other deficiencies. This reviewer found 5 items which permitted two correct responses (items 11, 14, 23, 27 and 35). One of these is the sample test item reproduced in this review. The keyed answer is D ("All convertibles are cars"). On the assumption that the subject class of each premise is non-empty, alternative B ("Some cars are convertibles") also follows from both premises together. The sample problem and an Euler circle diagram of the argument appears below:

All automobiles are cars.
All convertibles are automobiles.

Therefore:

A. Some automobiles are not convertibles.
B. Some cars are convertibles.
C. All cars are automobiles.
D. All convertibles are cars.



The test authors do not state that subject classes should be considered non-empty. They say only that test items are intended to be "true and factual in the realistic sense" (manual, p. 1), which might suggest that subject classes should be taken to be non-empty. In the absence of any explicit statement to this effect, however, the question of whether the five test items identified in this review permit more than a single correct answer must remain unresolved.

Further, the time limits suggested for this test seem far too short. This reviewer suspects that the test is highly speeded for a majority of examinees. We should therefore suspect that estimates of reliability might be inflated. No discussion of the speed and power aspects of the test is provided in the manual. In the opinion of this reviewer, the reliability estimates presented in the test manual must be viewed with caution, given that the method of their computation is left unspecified, and that the estimates themselves are likely to be inflated if they have been computed by Kuder-Richardson or split-half techniques.

The norms provided could probably be updated. *The Fifth Mental Measurements Yearbook (1959)* contains two reviews of the test, both of which identify the same norms currently appearing in the test manual. Such norms should probably be re-examined before being used as a basis for evaluation of examinee performance on the test. The lack of information about when such norms were obtained represents an unfortunate omission from the test manual.

Finally, the validity argument presented in the manual is unconvincing. The only relevant evidence presented for the validity of the test is a single factor-analytic study, whose reference is not cited. Correlations

Comments (continued):

presented for a similar form of this test provide only suggestive support, at best, for the validity argument. And the correlations reported in the manual permit no clear interpretation since no theoretical system is provided within which they might be understood as contributing to the construct validity of the instrument. This reviewer sees only a tenuous connection between the correlations reported for scores on Thurstone's four tests and Air Force Officer Candidate School criteria of achievement, and the construct validity of this test. It is incumbent upon the test authors at least to make clear the purposes for which such correlations are included in the test manual.

In general, the Logical Reasoning test is probably best considered as a measure of class reasoning ability whose validity and reliability still needs to be investigated.

Additional information:

Buros, O. K. The fifth mental measurements yearbook (1959). Reviews appear on pp. 694-695.

Hills, J. R. Factor-analyzed abilities and success in college mathematics. Educational and Psychological Measurement, 1957, 17, 615-622.

Hills, J. R. The relationship between certain factor-analyzed abilities and success in mathematics. Doctoral dissertation, University of Southern California, 1955.

Test citation:

Hertzka, A. F., & Guilford, J. P. Logical Reasoning. Orange, California: Sheridan Psychological Services, Incorporated, 1955.

Author(s): Educational Testing Service
Year of publication: 1950
Most recent edition: 1950

Description:

The Logical Reasoning Test was developed as an evaluation instrument in the Eight-Year Study of the Progressive Education Association. The test was constructed in two forms, Form A and Form B, which follow the same format. According to the test manual, this instrument "...represents an attempt to measure the ability of students to apply certain important principles of logical reasoning in appropriate situations" (p. 3). Specifically, the manual provides the following statements of the logical principles emphasized in the test:

(1)	definition:	Crucial words and phrases must be precisely defined, and a changed definition may produce a changed conclusion although the argument from each definition is logical.
(2)	indirect argument:	The validity of an indirect argument depends upon whether all of the possibilities have been considered. If there are three, and only three, possibilities, and one of them must happen, then, if two of the possibilities are shown to be, in fact, impossible, the third must happen. The conditions necessary for the logical use of indirect argument are seldom fulfilled in practice.
(3)	argumentum ad hominem:	An attack upon certain aspects of a person or institution, even though justified, is not sufficient to prove the lack of all merit in that person or institution. This covers the common use of ridicule, attack of motives, etc.
(4)	if-then:	If one accepts certain premises, then one must accept the conclusion which follows from these premises. The if-then principle is a necessary part of our method of criticizing generalizations, questioning assumptions, etc.

(pp. 3-4)

Each test form consists of four problems; one problem is intended to assess mastery of a single principle.

Each test problem consists of a passage which presents some line of reasoning or evidence which examinees must evaluate. Examinees are asked to assume the truth of statements given in the passage, and to select from

Description (continued):

three alternatives provided the conclusion which is justified by the information given in the passage. The possible conclusions for each passage follow this general form: (1) the argument presented in the passage establishes its intended conclusion; (2) the argument presented in the passage fails to establish its intended conclusion; and (3) additional information is needed in order to determine if the argument presented in the passage establishes its intended conclusion. The three possible conclusions are followed by a series of 12 statements. For each statement, the examinee must decide if: (1) it explains why the chosen conclusion is logical; (2) it does not explain why the chosen conclusion is logical; or (3) it leaves him undecided about whether it explains why the chosen conclusion was logical. The 12 statements to be evaluated are of three types, according to the description in the test manual: (1) general or abstract statements of the logical principle relevant to the test problem; (2) specific statements of the logical principle relevant to the test problem; and (3) general or specific statements of logical principles not relevant to the test problem, appeals to authority, statements of personal opinion and otherwise irrelevant statements. The manual points out that because of the small number of test problems, the responses to the 12 statements following the passage in each test problem receive considerable emphasis when examinee performance is evaluated. A time limit of 40 minutes is suggested for the test, although the manual states that the limit may be extended if necessary. A sample problem is provided for each test form.

Scoring is complicated and requires the use of special scoring keys provided with the test materials. The manual discusses the interpretation of test scores under the assumption that an individual examinee is administered both forms of the test. The reason for this treatment is suggested by the following statement in the test manual: "Since there are only a few questions in each form of the test, a score on one form alone will not be a very meaningful or reliable measure when the intention is to evaluate the abilities of individual students" (p. 7). Eleven scores are computed for each examinee: (1) number of correct conclusions chosen; (2) number of incorrect conclusions chosen; (3) number of relevant statements judged relevant; (4) number of relevant statements judged irrelevant; (5) number of irrelevant statements judged relevant; (6) number of irrelevant statements judged irrelevant; (7) number of "uncertain" responses and omitted items; (8) accuracy score on definition principle; (9) accuracy score on indirect argument principle; (10) accuracy score on argumentum ad hominem principle; and (11) accuracy score on if-then principle. Accuracy scores in each case are computed as the sum (over test forms) of the statements marked correctly for the two problems based upon each principle (maximum possible score: 24 for reach principle). The multiplicity of scores generated by this test is intended to distinguish examinees who arrived at a correct response by chance from those who have mastered a particular logical principle.

Description (continued):

The manual which accompanies the test contains clear directions for test administration and scoring; discussions of the logical principles addressed by the test; interpretation of test scores; and limited reliability information. No validity information is provided. In addition, further information regarding the development of this instrument can be found in Smith and Tyler (1942), referenced below. No norms are given in the manual, although means, standard deviations, standard errors, and standard errors of the difference for all eleven scores are presented for a group of 351 high school students in grades 10-12. Of this group, 292 students were completing grade 10, 28 were in grade 11, and the remainder were in grade 12. No analysis by grade level is given, and no additional information on the composition of the group is provided. These data suggest that the test should be appropriate for students in grades 10-12 but no specific recommendation is made in the test manual. However, the manual does state that the test is recommended "...only for classes where conscious attention has been directed toward logical reasoning" (p. 3).

Sample test item (Form B):

The following paragraphs are adapted from a newspaper article of January 24, 1940:

> George N. Peek, former president of the Export-Import Bank, appeared today before the House Ways and Means Committee, which is considering an administration request to continue the Reciprocal Trade Act. This Act has permitted the State Department to make trade agreements with foreign countries without consulting Congress.
>
> Mr. Peek declared that if trade agreements are treaties, the Constitution clearly requires their approval by the Senate; if they are tariff legislation, the Constitution positively requires that tariff legislation originate in the House.

Directions: Examine the conclusions given below. If the statements made by Mr. Peek are true, which one of the conclusions do you think is justified?

Conclusions

X. The above statements made by Mr. Peek proved that the Reciprocal Trade Act is contrary to specific requirements of the Constitution.

Y. The above statements made by Mr. Peek did not prove that the Reciprocal Trade Act is contrary to specific requirements of the Constitution.

Sample test item (Form B) (continued):

Z. Further information is needed to decide whether or not the above statements made by Mr. Peek proved that the Reciprocal Trade Act is contrary to specific requirements of the Constitution.

	A:	Statements which explain why your chosen conclusion is logical.
Mark in column	B:	Statements which do not explain why your chosen conclusion is logical.
	C:	Statements about which you are unable to decide.

Statements

1. Mr. Peek made a sound indirect argument from which the conclusion follows logically.

2. The conclusion is not necessarily true, but if one accepts the assumptions on which the argument is based, then to be logical one must accept the conclusion which follows from them.

3. It is quite possible that trade agreements are neither treaties nor tariff legislation.

etc.

Reliability and validity information:

No validity information is reported in the test manual. Reliability estimates for part scores were computed on the basis of administering both forms of the Logical Reasoning Test to the group of 351 students mentioned above. The manual does not identify the method by which reliability estimates were computed. Reliability estimates ranged from r = .23 (score on correct conclusions) to r = .88 (number of irrelevant statements judged irrelevant). Intercorrelations among accuracy scores on principles are also reported, and vary between r = .25 (definition and if-then) and r = .48 (indirect argument and ad hominem).

Availability:

The test is now out of print.

Comments:

The Logical Reasoning Test is an ambitious attempt to assess mastery of four logical reasoning principles and to provide a detailed appraisal of examinee performance in taking the test. The instrument, however, suffers from several serious deficiencies.

Examinees are asked, in each test problem, to choose the conclusion they feel is justified by the information presented. For those problems in which none of the three alternative conclusions follows deductively from the information presented, two responses seem equally appropriate. An examinee may correctly decide that the line of argument presented in the passage fails to establish its intended conclusion, and select the alternative which expresses this idea. However, an examinee could also reasonably judge that additional information is needed in order to determine if the argument presented in the passage establishes its intended conclusion. Each of these conclusions seems justified by the information given in the passage when such information is insufficient to establish the intended conclusion deductively. According to the answer key, however, judging that additional information is required is never the correct alternative. The test developers apparently intended examinees to consider only the logical form of the argument in evaluating whether the argument established its intended conclusion, although this restriction is not specified in either the test directions or test manual.

Further, it seems odd that given the emphasis placed on relevance scores, the test manual fails to mention that the test might also assess the ability to judge the relevance of reasons offered in support of an argument. Instead, the manual states that the percentage of irrelevant statements judged as irrelevant "...is the best single index on the test of the student's ability to apply the four principles of logical reasoning" (p. 10). This index was presumably chosen because it is the most reliable ($r = .88$). However, since both test forms combined contain 60 statements keyed as irrelevant and only 36 statements keyed as relevant, we should expect the irrelevant-judged-irrelevant score to be somewhat more reliable than the relevant-judged-relevant score. The test manual presents no rationale for considering the irrelevant-judged-irrelevant score as the single best index of the ability to apply the four logical principles emphasized in the test.

The test is also somewhat misleading in its use of the technical term, "indirect argument". As it is defined in the test manual, an indirect argument is not an indirect proof. An indirect proof (also called reductio ad absurdum) is one in which the assumption of the falsity of the desired conclusion is shown to lead to absurd or contradictory results. However, in describing the development of this test, Smith and Tyler (1942) equate their "indirect argument" principle with reductio ad absurdum (p. 114). The two principles are clearly not equivalent. It also seems likely that examinees might be unclear about the intended meaning of "indirect argument"

Comments (continued):

and that some examinees might take "indirect argument" to mean "indirect proof". However, since the test is recommended "...only for classes where conscious attention has been directed toward logical reasoning," the developers perhaps felt that examinees would be familiar with this special use of "indirect argument".

The absence of validity information is a critical deficiency of the instrument. The test manual provides no evidence to suggest that the test is valid for the purposes for which it was designed.

Finally, the failure to describe the method(s) by which reliability estimates were calculated represents a serious omission. The reliability estimates were calculated on the basis of administering both test forms to a group of 351 students. The manual does not state that a single administration was used, although in the judgment of this reviewer, this seems likely. On the assumption that the test forms are parallel, an examinee's scores on one form might have been correlated with his scores on the other form, and the Spearman-Brown formula used to obtain the reliability estimates for a test of double length (i.e., consistent with the recommendation that both forms of the Logical Reasoning Test be administered). Applying the Spearman-Brown formula to the reliability data given in the manual, each form of the Logical Reasoning Test yields reliability estimates which range from $r = .14$ (correct conclusions score) to $r = .45$ (irrelevant statements judged as irrelevant score). These reliability estimates suggest that a single form of the Logical Reasoning Test fails to provide a consistent appraisal of examinees' logical abilities. The Logical Reasoning Test, in order to be useful, should probably be thought of as the combination of both test forms. Even then, however, several of the reliability estimates are quite low, and the validity of the (combined) test still needs to be established. In its present form, the Logical Reasoning Test seems far from a finished product.

Additional information:

Buros, O. K. Fourth mental measurements yearbook (1953). Review appears on pp. 615-617.

Ennis, R. H., & MacMillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Smith, E. R., & Tyler, R. W. Appraising and recording student progress. Adventure in American Education, Volume III. New York: Harper and Brothers, 1942.

Test citation:

Educational Testing Service-Logical Reasoning Test. Princeton, New Jersey: Cooperative Test Division, Educational Testing Service, 1950.

Author(s): Robert H. Ennis
Year of publication: unpublished
Most recent edition: 1956

Description:

The Moorburg Letter Test is designed to assess students' reasoning abilities in the context of a letter-to-the-editor of a fictitious newspaper. The author of the letter is arguing for prohibition of parking on Moorburg city streets between the hours of 2 a.m. and 6 a.m. The reasoning offered in support of this position includes several formal and informal fallacies, as well as some legitimate support for the writer's position. The following errors in reasoning are included in the line of argument presented in the letter:

(1) some-to-all fallacy
(2) shift in meaning
(3) irrelevant material offered as evidence
(4) circularity
(5) affirming the consequent (particular conversion)
(6) weakness of evidence offered in support of position
(7) inadequate sampling
(8) post hoc fallacy
(9) winning argument by definition
(10) stipulated definition of word making term meaningless

The test consists of a total of nine 'items', divided as follows: (a) recognition of the some-to-all fallacy in the overall line of reasoning given by the writer of the letter constitutes one 'item'; (b) six paragraphs illustrate one or more fallacies of reasoning; (c) two paragraphs present legitimate support for the writer's position. The paragraphs of the letter are both numbered and set off from each other by spacing for reference. Examinees are given credit for identifying fallacious reasoning, justifying their identification of particular reasoning as fallacious, and recognizing the legitimate support given for the writer's position. The test is given in essay format.

Scoring is somewhat complicated, and requires the use of a special key. For each item, the possible range of scores is from minus 1 to plus 3. Scoring is done by the following system for each item:

score	interpretation
-1	error in identification of fallacious reasoning or in justification offered
0	omitted item
+1	correct identification of reasoning (either identification of fallacy committed, or identification of acceptable reasoning)

Description (continued):

score	interpretation
+2	correct identification of reasoning, but with incomplete justification offered
+3	correct identification of reasoning and adequate justification provided

Total test score is computed as the sum of the item scores; this formula allows a range of possible scores between minus 9 and plus 27.

No manual is provided with the test; hence, no recommendation is made concerning the age groups for which the test might be appropriate. However, the test has been administered to grade 10 and 11 students as part of the research accompanying the development of the instrument. No time limit is specified for completion. Test directions are clear and straightforward.

Sample test item:

7. The opponents of my suggestion have maintained that conditions are safe enough now. These people don't know what "safe" really means. *Conditions are not safe if there's ever a conceivable chance for an accident*. Certainly conditions are not safe the way they are now.

[emphasis in original]

Reliability and validity information:

Some reliability and validity data were obtained in the process of developing this instrument. Students in grades 10-11 from three schools in Illinois were used for the purposes of this study. The current form of the Moorburg Letter Test was administered to 96 students, from which 34 pairs of students matched on subjects taken in school, grade point average, class and sex (where possible) were selected. One member of each pair had been randomly selected from classes which had been exposed to instruction in some reasoning skills for varying periods of time. Instruction was provided by regular classroom teachers. The other member of each pair was matched on the variables noted above, and was drawn from classes which had not received such instruction. When mean scores on the Moorburg Letter Test for the instructed vs. non-instructed groups of students were compared, the instructed group achieved a significantly higher mean score ($p < .05$). Examinees' scores were determined independently by two graders who were unaware of the status of each student (instructed or non-instructed). In cases of scoring disagreement, the scoring was reviewed. If the disagreement

Reliability and validity information (continued):

persisted after review, the examinee's score was determined as the average of the two scores. An inter-rater reliability coefficient was computed for the scores before review, and found to be r = .88.

No norms are presented.

Availability:

The test is unpublished, and is currently being reviewed for possible future development by the Illinois Critical Thinking Project.

Comments:

The Moorburg Letter Test is still in its experimental edition. Despite this fact, however, the instrument has several promising features. Although this instrument is not a general measure of critical thinking ability, it does attempt to assess individuals' reasoning abilities in a context.

The test requires examinees to identify a number of informal fallacies which are often overlooked in tests of reasoning ability. In addition, examinees are presented with a line of argument in context, rather than in isolation, and are asked to evaluate the reasoning used to support the conclusion. This situation is more representative of the types of circumstances likely to require critical reasoning (especially since examinees must recognize legitimate support when it is given), and the test might therefore provide a more realistic appraisal of an individual's reasoning abilities.

The essay format is both a strength and a weakness of this test. Scoring requires greater time for essay tests, and may be subject to unreliability due to differences in raters. Machine scoring is impossible. Further, the scorers must be familiar with the material upon which the test is based. Straightforward comparison of examinee responses with a set of keyed answers is impossible, and scorers must depend upon their own judgment in awarding credit for items. The fairly high inter-rater reliability coefficient obtained in the development of this test suggests that reasonable agreement among scorers is a practical goal, however.

On the other hand, essay tests may allow more methodical students an opportunity to perform to their best capability. The requirement that examinees justify their responses seems best suited to non-multiple-choice tests, since even multiple-choice tests which required examinees to select appropriate justification for their responses from the alternatives provided, could not establish that examinees themselves could provide such

Comments (continued):

justification. In short, the Moorburg Letter Test might come closer to requiring examinees to demonstrate critical reasoning, rather than to identify instances of it.

Probably the most serious problem confronting the test at this time is the lack of information concerning it. In the opinion of this reviewer, the instrument merits further development. Reliability and validity investigations might suggest profitable lines for future research. Of course, preparation of a manual to accompany the test is necessary if the Moorburg Letter Test is to be considered a finished product.

Test citation:

Ennis, R. H. Moorburg Letter Test. Unpublished test. Urbana, Illinois: Project for the Improvement of Thinking, University of Illinois, 1956.

Author(s): Euclid Elementary teachers
Year of publication: 1944
Most recent edition: 1946

Description:

The Ohio Thinking Checkup was constructed by a cooperative effort of the College of Education, Ohio State University and the Euclid Elementary teachers. According to the test manual, the instrument should "...give some evidence as to a child's ability to think with discrimination [and] also provide some idea as to the child's sense of values" (p. 26). The test consists of a series of 10 short stories, each followed by 8 statements. Examinees must judge, on the basis of the information given in the story, whether each statement is true, false or that he cannot tell whether the statement is true or false.

According to an article by Louis Raths in the Educational Research Bulletin (referenced below), the instrument is designed to reveal the following errors in thinking:

1. Interpreting through personal judgment which is external to the data
2. Evading the issue by name-calling or ridicule
3. Leaning on authority
4. Believing in a superstition
5. Generalizing from insufficient evidence; that is, what is true in one or a few cases is true in all
6. Rationalizing or misinterpreting data
7. Calling either-or statements "true"
8. Calling if-then statements "true"
9. Leaning on school loyalty

(p. 73). Raths notes that each error in reasoning was represented twice in the course of the test (p. 73).

Four scores are computed for this test, which are intended to reveal the following aspects of an individual's reasoning ability:

(1) degree of accuracy in responding to factual data (percentage of 'true' responses to items keyed as 'true')

(2) error in thinking from facts (percentage of 'false' responses to items keyed as 'true')

(3) tendency to go beyond the data (percentage of items marked 'true' or 'false' when correct response is keyed as '?')

(4) caution (percentage of items marked '?' when correct response is keyed as 'true' or 'false')

Description (continued):

Scores are computed on the basis of the number of items attempted within whatever time limit is specified by the test administrator. The manual includes a brief discussion of score interpretation, and presents hypothetical examples of possible score distributions.

The test manual makes no recommendation concerning the age groups for which the test is appropriate. However, Louis Raths indicates that the test is intended for use in grades 4-6, but may also be used for junior high school students. No time limit is specified for completion of the test, and the manual specifically states that "...there is no intention of making this a test of reading speed or comprehension" (p. 27). Raths suggests that 30-40 minutes should be adequate for most students (p. 84).

Test directions are generally clear and the language used seems appropriate for grades 4-6 students, although it may be somewhat too simple for students in junior high school. In addition, Raths claims that the stories used in the test were selected as representative of situations in which grades 4-6 students were likely to make generalizations, and were also within the experience of the students for whom the test was designed (pp. 73-74).

The manual provides no reliability or validity information; teachers are advised to compute their own norms.

Sample test item:

[The following directions appear on the first page of the test booklet. Directions are not repeated at any point during the test. The test booklet states that the directions are to be read aloud.]

Sometimes when we read something we think about it. Sometimes when we hear something we think about that, too. We ask ourselves, "Is that true?" Often we say, "That isn't true." At other times we can't make up our minds. We are uncertain.

We have some stories for you to read. After each story you will find some sentences. We want to know what you think about each sentence. You don't have to believe the whole story. You should think about the story and think about each sentence. If you think the sentence is true, draw a circle around the "T"; if you think the sentence is false, draw a circle around the "F"; and if you cannot tell whether it is true or false, draw a circle around the "?". If you cannot read it or understand it, raise your hand and your teacher will help you to read the sentence. Make only one (1) circle for each sentence.

[test, p. 1; emphasis in original]

Sample test item (continued):

Sam was walking to the store one day when a black cat crossed his path. Harry saw this from across the street and shouted, "You're going to have bad luck." That night Sam lost a quarter which his mother had given him for a movie. The next day Harry said, "See, I told you that black cat would bring you bad luck."

1. A black cat crossed Sam's path and that night he lost a quarter that his mother had given him for a movie . . . . T F ?
2. Harry was glad that Sam lost the quarter . . . . . . T F ?
3. If a black cat crosses your path, then you are sure to have bad luck. . . . . T F ?

[etc.]

Reliability and validity information:

Although the test manual does not report any reliability or validity data, Louis Raths presents some relevant information in his Educational Research Bulletin article. Some crude evidence of the test's validity was obtained by comparing teacher rankings of students' reasoning abilities and performance on an earlier version of the Ohio Thinking Checkup. Students ranked as superior thinkers and those ranked as the poorest thinkers by their teachers were used for comparison purposes. The number of students sampled is not stated, although Raths notes that "several hundred" students in the Euclid schools and students in "four other schools in the state" were administered this form of the test (p. 75). Raths states that "...there was relatively high agreement between the rankings of the teachers and the rankings of students' scores on the test" (p. 75). The means of determining the extent of "agreement" between teacher ranking and test performance is not stated. Raths also cautions that "...no conclusions may be drawn at this writing concerning the ultimate validity of the instrument" (p. 75).

A measure of the reliability of an earlier form of this test was also computed. Each examinee's performance on one opportunity to commit a particular error in reasoning was correlated with his performance on the second opportunity to commit that error in reasoning. Reliability estimates ranged from r = .89 (grade 4 students) to r = .93 (grade 6 students). Raths does not state whether the reliability estimates reported in his article apply to the test as a whole or to a 'half test'.

The test manual advises teachers to compute their own norms.

Availability:

The test is currently out of print and was formerly available from the Ohio State Department of Education.

Comments:

The Ohio Thinking Checkup appears to be an easily-administered, easily-scored instrument which may be of some value in assessing the reasoning abilities of younger students. The appeal of the instrument is in the considerable information potentially provided by the measure, and the fact that the scores obtained on the test permit a relatively clear interpretation. Unfortunately, however, the test suffers from several problems which prevent it from achieving its full potential.

No reliability or validity data for this form of the test is provided in either the manual or in the article by Louis Raths. The absence of such information suggests that scores on the instrument must be interpreted cautiously, and that individuals desiring to use the test must rely completely on their own judgments of its appropriateness.

Although the test manual mentions that the test may reveal an examinee's values, there is no discussion of how the instrument might provide such a measure, or how whatever information about individual values the test provides should be interpreted. And because many of the statements which follow each story confound sources of error, it is unclear whether the keyed answers are the only correct responses. For example, an examinee who decided that "All people who speak English well are always very bright" (story nine) is false might do so because he rejected the values contained in the statement. However, an examinee might also mark the statement '?' since no information is provided in the story which would suggest that the statement was true or false.

A further difficulty with the test is the problem of determining which statements are intended to test for particular errors in reasoning. For example, statement 3 in the sample test item reproduced here might be intended to test for an examinee's tendency to believe in a superstition. It might also be intended to test for an examinee's tendency to call if-then statements 'true'. Both of these tendencies are identified as errors in reasoning by Raths. A discussion of the errors in reasoning purportedly represented by each statement would be a helpful addition to the manual.

Finally, this reviewer feels that the errors in reasoning identified by Raths require elaboration. If-then and either-or statements can be true or false; why should calling if-then statements of either-or statements 'true' necessarily represent an error of reasoning, (see points 7 and 8 under 'Description')? A clearer, more detailed account of the errors in

Comments (continued):

reasoning purportedly identified by this test would contribute greatly to an evaluation of the instrument's validity. Overall, the test shows some promise, and may be capable of providing a reasonably accurate appraisal of the reasoning abilities of younger students. The test suffers, however, from incompleteness in a number of important areas. In its present form, the Ohio Thinking Checkup is probably best considered an unfinished (but promising) product.

Additional information:

Ennis, R. H., & MacMillan, C. J. B. Critical thinking reviews (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Raths, L. E. A thinking test. Educational Research Bulletin, 1944, 23, 72-75, 84.

Test citation:

Euclid Elementary teachers. The Ohio Thinking Checkup for Intermediate Grades. Columbus, Ohio: College of Education, Ohio State University, 1944.

Author(s): G. H. Hyram and W. A. Pearson
Year of publication: unpublished
Most recent edition: 1958

Description:

The Pearson-Hyram Test of Critical Thinking was originally devised as an evaluative instrument in Hyram's doctoral work and was used in a subsequent study (referenced below). W. A. Pearson performed the item analysis procedures which resulted in the selection of the items which comprise the current form of the test. No manual accompanies the test, and it is therefore impossible to be certain what traits or abilities the instrument is intended to measure. However, Hyram (1957) identified seven components of logical thinking in his research. Since an earlier form of this measure was used to compare instructed and uninstructed groups, it seems likely that the test performance was intended to reflect abilities in the seven areas listed in the article, and that the current form of the Pearson-Hyram Test of Critical Thinking also assesses at least some of the same abilities. The components of logical thinking included in the instruction given by Hyram (1957) were: (1) nature of thinking; (2) tools of thinking; (3) nature of definition; (4) nature of eductive inference; (5) nature of deductive inference; (6) nature of experimentation; and (7) common errors in reasoning.

The test consists of 10 sections, and a total of 112 questions. Items 1-38 comprise Part I of the test which is entitled, 'Grasping Implications'. The first 30 items present students with two premises. In some items, examinees must decide which of four alternative conclusions follows necessarily or that none of the alternative conclusions follows necessarily. In other items, examinees must decide which of four alternative conclusions cannot follow logically from the premises provided, or that any of the alternative conclusions could be correct. Questions 31-34 present a short paragraph and one or two facts given in the paragraph. From the fact(s) provided, examinees must judge which of three alternative conclusions follows logically, or that none of the alternative conclusions follows logically. Questions 35-38 require examinees to label appropriately the circles in an Euler circle diagram of a class reasoning argument.

Part II (items 39-56) is entitled, 'Finding Errors in Thinking'. Examinees are presented with two 'given facts' and a conclusion, and are asked questions about each. Examinees must respond 'yes' or 'no' to each question. The final two items (questions 51-56) deal with the nature of good thinking and obstacles to good thinking; examinees respond 'yes' if they believe a given statement to be true, and 'no' if they do not.

Part III (items 57-68) is entitled, 'Grasping Implications from Hypothetical (Assumed) Situations'. Examinees are given conditional arguments in symbolic form and asked to judge which (if any) of three alternative conclusions follows necessarily, or that none of the alternatives follows necessarily. Part IV (items 69-72) is labelled 'Discovering Hidden Assumptions'. For short arguments, examinees must choose which of three alternatives represents a missing, yet understood part of the argument, or that they cannot tell what the missing part is. Part V ('Dilemmas') consists of items 73-76, which present examinees with arguments cast in the form of a

Description (continued):

dilemma. Examinees must decide which of three alternative conclusions must follow from the information given, or that none of the alternatives necessarily follow. Item 76 requires examinees to recognize that the failure to show that only two alternatives exist is the fallacy of reasoning contained in items 73-75. Part VI (items 77-94) is labelled 'Discovering Rules' and is divided into two sections. In the first section (items 77-88), examinees must choose the next member in a series of letters or numbers from four alternatives, or judge that the correct answer is not among those provided. Each question in the second section (items 89-94) presents examinees with four groups of four letters. Examinees must determine which group of letters differs from the other three.

Part VII ('Building a Theory') consists of items 95-98, which are taken directly from Sir Cyril Burt's Graded Reasoning Test. Examinees are given 'detective stories' and are asked to infer conclusions based upon the information provided. Part VIII ('Making Generalizations') is divided into two sections. In the first section (items 99-102), examinees must choose which of three alternative procedures should be followed in obtaining supporting evidence for a theory, or judge that none of the alternatives is appropriate. Section two (items 103-104) asks examinees to choose which of three alternative generalizations best explains the content of a given passage, or that the facts provided in the passage are insufficient to serve as a basis for any conclusion. Part IX ('Making Analogies'--labelled as Part X in the test) is comprised of items 105-108. A short passage is presented describing similarities between two or more things. Examinees must respond either 'yes' or'no' to two questions for each passage: (1) is a given inference, based upon the similarities described in the passage, reasonably sound? and (2) is the inference established by the information given? Part X (identified as Part XI in the test; 'Identifying Errors in Thinking') includes items 109-112. Examinees are given short arguments and must judge which of two reasons accounts for the deficiency of the argument.

Because no manual is provided with the test, no details of test administration, scoring, score interpretation, reliability and validity information or norms are available. Inspection of the test suggests that administration time will be lengthy. Hyram (1957) does not identify the grade levels of the subjects used in his research; hence, there is no way of knowing for whom the test is intended to be appropriate.

Test directions are generally clear, although occasionally confusing (see 'Comments', below). Individual items sometimes appear to be awkwardly worded.

Sample test item:

For the following items, determine the listed conclusions that ARE NOT TRUE. Blacken the letter of the conclusions that MUST NOT BE TRUE in the spaces provided on the answer sheet.

10. If every dog is an animal, then it is not true that--

    1. some dog is an animal
    2. no dog is an animal
    3. some animal is a dog
    4. a dog is a cat
    5. any of the above conclusions is correct

[emphasis, capitalization in original]

Reliability and validity information:

No reliability or validity information is currently available for this test.

Availability:

The test was apparently never published.

Comments:

The Pearson-Hyram Test of Critical Thinking suffers from a number of defects. In the sample test item reproduced in this review, the directions to the examinee are confusing. One statement directs examinees to select conclusions which are not true; the other statement requires examinees to determine which conclusions must not be true. The confusion here is between selecting alternatives on the basis of their factual content (for this example, alternatives 2, 4 and 5 are factually false) and selecting alternatives on the basis of their logical status with respect to the given premises (alternative 2 clearly cannot be true if the premises are accepted, alternatives 4 and 5 do not follow logically from the premises, but neither do they contradict the premises). It is important to note here that the directions do not specify that alternative conclusions are to be evaluated on the basis of whether they follow logically from the premises. And alternative 5 (identical for all items of this type) is always a correct answer, since the items in this section all contain at least one conclusion which contradicts the premises or one conclusion which is factually incorrect. In addition, alternative 1 will follow logically from the premises and be

Comments (continued):

factually true if the subject class is assumed to be non-empty. However, alternative 1 will not follow logically from the premises and will be factually false if this assumption is not made. And it should be noted that the fifth alternative may be interpreted as a possible completion for the stem "...it is not true that--" or as an independent response. No distinction is drawn between these interpretations in the test; consequently, the fifth alternative remains a source of confusion which may affect examinees' responses. A different problem with test directions occurs in another section of Part I. Examinees are asked to read an argument and a conclusion which follows from it; the conclusion, however, is not provided.

Numerous test sections contain only a limited number of items. These sections include items 35-38 (Euler circles); items 51-56 (nature of sound thinking); items 69-72 (Part IV--'Discovering Hidden Assumptions'); items 73-76 (Part V--'Dilemmas'); items 95-98 (Part VII--'Building a Theory'); items 99-102 (supporting a theory); items 103-104 (judging explanatory power of generalizations); items 105-108 (Part IX--'Making Analogies') and items 109-112 (Part X--'Identifying Errors in Thinking'). It seems unlikely that the small number of items in each of these sections could provide reliable measures of whatever abilities are tested.

The strength of the Pearson-Hyram Test of Critical Thinking lies in Part I and Part III, which together may provide some assessment of deductive reasoning ability. Part I deals with class reasoning principles, Part III deals with conditional reasoning principles, and includes one principle of alternation logic. Unfortunately, however, only a limited number of logical principles are represented in the test items. Inferences about an examinee's ability to reason deductively based on performance on these two test sections should therefore be made cautiously.

Overall, the Pearson-Hyram Test of Critical Thinking is far from a finished product, although it does show some promise. In its present form, and in the absence of any information about test validity, reliability and construction, it is probably best viewed as a source of ideas for the development of other critical thinking measures.

Additional information:

Hyram, G. H. An experimental study in developing critical thinking in children. Journal of Experimental Education, 1957, 26, 125-132.

Hyram, G. H. An experimental study of classroom procedures for improving the general reasoning ability of upper grade pupils. Doctoral dissertation, St. Louis University, 1956.

Additional information (continued):

Pearson, W. A. An item analysis and revision of the Hyram test of critical thinking. Master's thesis, St. Louis University, 1958.

Test citation:

Hyram, G. H., & Pearson, W. A. Pearson-Hyram Test of Critical Thinking. Unpublished test. St. Louis: St. Louis University, 1958.

Author(s): Instructional Objectives Exchange
Year of publication: 1971
Most recent edition: 1971

Description:

Recognizing Reliable Observations is a criterion-referenced measure intended to assess examinees' proficiency in judging the relative reliability of observation statements. The instrument is based upon criteria for judging the reliability of observation statements identified by Ennis (1962). The test consists of 10 items. Each item presents three statements, and examinees must indicate which of the three statements "...is the one in which [he or she] would have the most confidence" (manual, p. 61). Scoring is by the 'rights only' method, and a total test score is computed as the sum of correct responses. A time limit of 10-15 minutes is suggested for completion. A sample problem is provided at the beginning of the instrument.

Sample test item:

2. a) Because the child's head felt hot, his mother said he had a fever.

   b) Because the child acted cranky and looked flushed, his mother said he had a fever.

   c) Because the child's temperature read 103 on a thermometer, his mother said he had a fever.

Reliability and validity information:

No reliability or validity information is provided for this measure, beyond the suggestion made in the manual that the items in the test were judged appropriate (i.e., congruent with the behavioral objective) by at least one subject matter expert. However, the meanings of test reliability and test validity have not yet been adequately specified for criterion-referenced tests. The lack of reliability and validity data probably reflects this situation, at least in part.

Availability:

The test is currently in print and is available from the Instructional Objectives Exchange, P.O. Box 24095, Los Angeles, California, 90024.

Comments:

This measure is easily administered and scored. A number of test items are clearly based upon the criteria developed by Ennis (1962), which should

Comments (continued):

aid in identifying specific difficulties of examinees in judging the reliability of observation statements. However, the instrument suffers from two serious problems.

Many of the statements provided in the test items are not observation statements. The sample test item reproduced here, for example, presents three (identical) inferences, each based upon a different observation. Examinees are not evaluating observation statements except insofar as the various observations provide differing degrees of justification for the inference based upon them ("...his mother said he had a fever"). No observation statements are made, however; the only statement given in the sample test item appears to this reviewer to be an inference, rather than an observation. Other items also suffer from similar problems. Item 8, for example, describes three measuring techniques without presenting any statement of observation; item 10 gives a description of three methods of timing a race, but includes no observation statements. These sorts of difficulties with the test items are likely to make interpretation of test performance problematic. At the very least, it is not clear that such items as those identified above should be considered legitimate members of the item population from which test items were (theoretically) drawn.

The other significant problem of the test is the limited number of items (10). Not all of the criteria identified by Ennis (1962) are represented in the test items; neither is any description provided of which criteria are tested for by each item. In view of this situation, it seems likely that a score on this test cannot be given a clear interpretation.

In general, Recognizing Reliable Observations appears to be a measure still in its beginning stages of development. In the opinion of this reviewer, the instrument requires considerably more work if it is to be used as a measure (criterion-referenced or otherwise) of ability to judge the reliability of observation statements.

Additional information:

Ennis, R. H. A concept of critical thinking. Harvard Educational Review, 1962, 32, 81-111.

Instructional Objectives Exchange. Recognizing Reliable Observations. Los Angeles, California: Instructional Objectives Exchange, 1971.

Test citation:

Instructional Objectives Exchange. Judgment: Deductive logic and assumption recognition. Los Angeles, California: Instructional Objectives Exchange, 1971.

Author(s): Patricia King and Karen Kitchener
Year of publication: unpublished
Most recent edition: 1977

Description:

The Reflective Judgment Interview was developed as part of King's doctoral research, and is based largely on the theory of intellectual and ethical development of college students offered by William Perry (1968). Perry considers both intellectual and ethical development together, in terms of nine stages or "positions." Each position is characterized by a particular epistemological perspective or "cognitive style." The first three positions are subsumed under the heading of "dualism"; the next two positions comprise the category of "relativism." The final category, "commitment in relativism", includes the final four positions. King argues in her dissertation that Perry's system is inadequate to deal with many questions of concern in the field of intellectual development. King therefore suggests an additional position ("probabilism") and proposes a new construct, reflective judgment, as adjuncts to Perry's theory. Reflective judgment is defined as the process of resolving problems or questions which do not have a single, logically valid answer (p. 16). A reflective judgment is distinguished from a formal or logical judgment by the fact that the latter deals with problems having "...one logically valid answer which could be verified by reference to logical rules of procedure, as, for example, in a syllogistic argument" (p. 15). In contrast, reflective judgments involve problems with "...more ambiguity...a variety of possible solutions or resolutions, and [which] may be resolved through a selection and evaluation of what the individual sees as relevant and important evidence" (p. 16).

The instrument consists of four dilemmas, and a standard set of probe questions. Each dilemma contains statements expressing two distinct viewpoints on a single topic. For each dilemma, examinees must respond to the series of probe questions. Their responses are then scored according to a set of rules developed by King in her research, and based partly on Perry's theory. These scoring rules are intended to identify the type of reasoning exhibited by the examinee along three dimensions: (1) cognitive complexity; (2) reasoning style; and (3) openness.

The measure is administered individually. A dilemma is read aloud to an examinee who follows along by reading the dilemma from a typed card. The probe questions are then asked. The test requires approximately 60 minutes to complete. Although no specific recommendations are made concerning the age groups for which the test might be appropriate, King studied students in high school through graduate school. King's dissertation also contains discussions of the development of the instrument, reliability and validity data, scoring procedures, and examples of reasoning characterizing each position.

Sample test item:

2. Some people believe that news stories represent unbiased, objective reporting of news events. Others say that there is no such thing as unbiased, objective reporting, and that even in reporting the facts, the news reporters project their own interpretations into what they write.

probe questions:

1. What do you think about these statements?
   (If the person does not take a stand on the issue, skip questions 2 through 5)

2. How did you come to hold that point of view?

....

(If the person does not take a stand on the issue)

2a. Could you ever say which was the better position? How/why not?
(If yes) What do you mean by "better?"

3a. How would you go about making a decision about this issue?

....

Reliability and validity information:

King presents some reliability and validity information for this measure. Two judges independently assigned 3 scores (corresponding to the 3 dimensions identified earlier) to the responses of each subject for each dilemma. Three groups of 20 subjects, ranging from high school to graduate students were used in the study. Inter-rater reliability for each of the four dilemmas across the total sample ranged from r = .86 (dilemma 1) to r = .90 (dilemma 2 and dilemma 5). Inter-rater reliability estimates for each student group across all dilemmas varied between r = .68 (high school students) and r = .94 (college students). Inter-rater agreement was operationally defined as a difference of 2 points or less between the sum of the three scores given by each rater.

Scores on the instrument also show a generally increasing trend with educational level, which supports the theoretical expectations of the study, and provides evidence for the construct validity of the instrument. For all subjects and all dilemmas, inter-dilemma correlations ranged from r = .84 (dilemmas 1 and 2) to r = .88 (dilemmas 3 and 4). Item (dilemma)-total score correlations for the group as a whole were also reported, and vary between r = .89 (dilemma 2) and r = .92 (dilemma 4). King also provides

Reliability and validity information (continued):

item-total score correlations for each student group. The reported correlations are corrected for the correlation of an item with itself. Coefficient alpha was also used as a measure of internal consistency, and was found to be .96 for all dilemmas across the total sample. Estimates for each student group ranged from alpha = .84 (graduate students) to alpha = .88 (high school students).

Availability:

The test is unpublished, but a copy of the instrument appears in King's doctoral dissertation (1977). This work also describes the study for which the instrument was developed.

Comments:

The Reflective Judgment Interview is an interesting attempt to classify an individual's cognitive "style" or approach to solving problems which generally require the evaluation of evidence rather than appeals to principles of deductive reasoning. The instrument seems potentially useful for diagnostic purposes, and represents a somewhat different approach to assessing intellectual development.

However, the measure has several drawbacks. The lengthy, individualized administration is likely to make using the instrument impractical in many cases. Further, scoring can be done only by a person familiar with the theoretical basis for the test, i.e., the work of William Perry and King's modifications of it. The validity and usefulness of scores on the Reflective Judgment Interview may well be determined by the adequacy of Perry's theory.

It also seems to this reviewer that the Reflective Judgment Interview does not measure some sort of cognitive ability per se. Rather, it appears to assess an individual's world view from both an epistemological and an ethical perspective. The instrument seems designed more to reveal an attitude than to identify specific skills and abilities which might be associated with the construct of critical thinking ability. This focus is not being criticized here; on the contrary, certain attitudes and dispositions (such as the willingness to suspend judgment and the willingness to consider alternative viewpoints) would seem to be prerequisites for developing and improving critical thinking ability. However, it is important to be clear about what a particular instrument is intended to assess. The Reflective Judgment Interview seems to this reviewer to be more affectively than cognitively oriented; prospective test-users should be aware that this measure may provide a useful addition to, but not a substitute for, other critical thinking measures focusing on cognitive skills and abilities.

Additional information:

King, P. M. The development of reflective judgment and formal operational thinking in adolescents and young adults. Doctoral dissertation, University of Minnesota, 1977.

Perry, W. G. Forms of intellectual and ethical development in the college years. New York: Holt, Rinehart and Winston, 1968.

Test citation:

King, P. M. Reflective Judgment Interview. Unpublished test. University of Minnesota,

Author(s): Loretta L. Molitor
Year of publication: unpublished
Most recent edition: 1972

Description:

The Science Process Skills Test is intended to assess the performance of examinees in grades 4-6 in three areas: (1) inference (to best explanation); (2) prediction; and (3) verification. The instrument is constructed in standard multiple-choice format, but the test items and item distractors are presented as illustrations. Molitor notes that this approach was used in order to minimize differences in reading ability which might distort scores on the instrument. In addition, test directions are read aloud to examinees, and the test administrator is instructed to keep all examinees together when taking the test; examinees must wait for directions to continue after completing each problem.

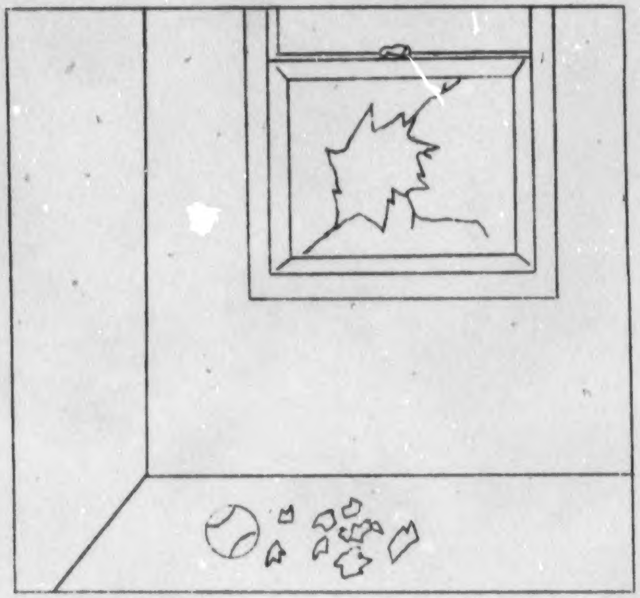
The instrument is composed of three subtests. The inference and verification subtests are administered together in a total of 9 items. Each item requires three responses, however; the first two responses for each item comprise the inference subtest, and the remaining response constitutes the verification subtest item. In these 9 items, examinees are presented with a picture and must make two inferences about the event suggested by the drawing. The first inference required is usually a causal inference; for example, a picture is presented of a broken window with a ball and pieces of glass lying on the floor. Examinees must choose from the three alternatives provided, the drawing which shows how the window was probably broken. Examinees then select which of three other drawings shows what information given in the original picture helped them to decide how the window was probably broken. Finally, examinees must choose which of the 6 previous drawings shows what they would have to see in order to verify their inference (in this case, to be sure how the window was broken).

The prediction subtest consists of 8 items. In each item, a series of pictures is presented, usually depicting some sort of progressive change. Examinees must select which of three alternative drawings shows the next event in the series, or judge that insufficient information is provided on which to base a prediction.

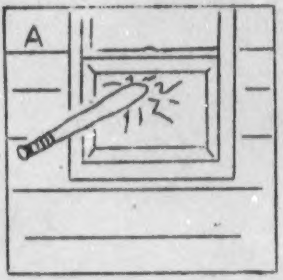
Each section of the test is preceded by a sample test item which is worked by students and test administrator together. Scoring of responses varies as a function of the type of item. Prediction subtest items are scored by the 'rights only' method. The inference items are scored plus 1 for two correct responses, otherwise, zero. Verification items are also scored by the 'rights only' method. Total test score is calculated as the sum of the part scores. No time limit is specified for the test, and test administrators are encouraged to adjust the time limits to the group being tested. Although no manual accompanies the test, Molitor (1972) serves this purpose. Separate instructions for test administration are also provided.

Sample test item (inference and verification):

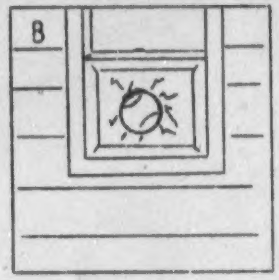
Look at the picture below.



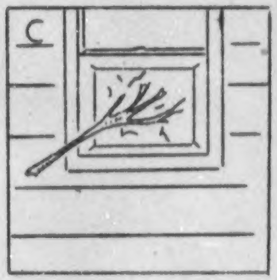
a) Put an X in the box below the picture that shows how the window most likely was broken.



☐

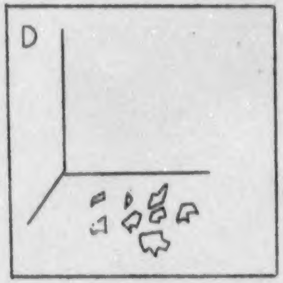


☐

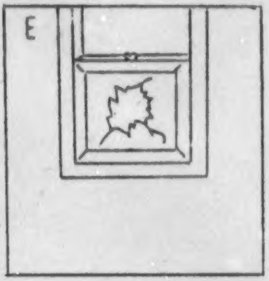


☐

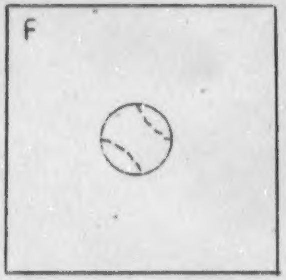
b) Put an X in the box below the picture that shows what you can see in the big picture that helped you decide how the window was most likely broken.



☐



☐



☐

c) Which one of all these six pictures shows what you would have to see in order to be sure how the window was broken?

Put the letter of that picture on this line. ________

Reliability and validity information:

Molitor (1972) presents considerable reliability and validity data for this instrument. She formulated both definitions of the skills involved in the three abilities sampled by the Science Process Skills Test, and behavioral criteria by which examinee behavior was evaluated as evidence of a particular skill or ability. A preliminary long form of the instrument, containing 17 items testing each skill, was submitted to a panel of three science educators. Those items judged unanimously to be consistent with both the definitions and behavioral criteria were retained in the revised long form of the instrument. The revised long form was administered to grade 6 children, and items of unacceptable difficulty and discriminating power were eliminated. The remaining items comprised the present form of the measure.

A sample of N = 120 students in grades 4-6 was used to test the hypothesis of no association between a correct item response and behavior consistent with the stipulated performance criteria, for each item at each grade level. Students were questioned about their reasons for selecting particular responses to items randomly sampled from the final form of the test. Student responses were summarized in a frequency table reflecting the four possible response patterns: (1) correct response, reason inconsistent with performance criteria; (2) correct response, reason consistent with performance criteria; (3) incorrect response, reason inconsistent with performance criteria; and (4) incorrect response, reason consistent with performance criteria. Items which showed a significant ($p < .025$) positive association between item response and behavior consistent with the performance criteria were judged to be valid, i.e., "...capable of illiciting [sic] behavior consistent with the stipulated performance criteria" (Molitor, 1972). Significant positive associations between item response and behavior consistent with the performance criteria were found for the majority of the inference and verification items; several prediction items failed to show any significant positive association.

Another validity study was performed using two groups of grade 4 students. One group consisted of subjects with five years of instructional experience with Science: A Process Approach. The remaining group was composed of similar subjects who had not been exposed to such instruction. The Science Process Skills Test was administered to both groups, and part and total test scores were compared. The instructed group was found to perform significantly better ($p < .01$) than the uninstructed group for all scores computed for this test.

The above groups of grade 4 students also provided data used in deriving estimates of part and total test score reliabilities. Kuder-Richardson 20 reliability estimates for these groups ranged from $r = .408$ (prediction subtest score) to $r = .714$ (verification subtest score). A sample of N = 96 grade 5 students was used as the basis for computing Kuder-Richardson 20 reliability estimates for part and total test scores at this grade level. Estimates ranged from $r = .260$ (prediction subtest score) to $r = .838$

Reliability and validity information (continued):

(verification subtest score). Similar calculations for a group of N = 100 grade 6 students yielded reliability estimates varying between r = .324 (prediction subtest score) and r = .793 (verification subtest score).

No norms are provided. However, Molitor (1972) presents total test and subtest means and standard deviations by grade level for the groups identified above.

Availability:

Although the test is unpublished, copies of the instrument may be obtained from the author at the following address: Physics Department, Science Education Group, Towson State College, Towson, Maryland, 21204.

Comments:

The Science Process Skills Test is noteworthy for several reasons: (1) the instrument is intended to assess the abilities of younger examinees on three important aspects of critical thinking ability; few (if any) currently available critical thinking tests are designed to assess these abilities at the elementary school level; (2) the instrument attempts to minimize the effects of differences in verbal ability which might influence test performance; (3) inference subtest items are scored on the basis of correct response and adequate justification, which potentially permits a more accurate interpretation of examinee performance. However, the instrument also suffers from several weaknesses.

The validity argument presented by Molitor (1972) seems to establish that some test items appearing on the present form of the instrument are invalid, as judged by her standard. These items should probably be eliminated or revised, in order to improve the validity of the measure as a whole.

Further, the reliability estimates presented for the prediction subtest are uniformly low at each grade level. These estimates suggest that the subtest items are not homogeneous, and the subtest could probably profit from further investigation. One explanation for the lack of item homogeneity might be that the items do not all measure the same ability. However, a test-retest reliability estimate for this subtest might reveal that the items provide consistent measurement of the ability or abilities sampled. Such a finding, interpreted in conjunction with the low Kuder-Richardson 20 reliability estimates, might suggest that the items differ from each other in the abilities sampled, but provide consistent measures of those abilities over time. Hence, the subtest may be reliable, but not measure only one ability. Molitor suggests eliminating the prediction subtest from the measure (1978; personal communcation).

Comments (continued):

An additional problem of the instrument concerns the instructions given to examinees on the verification items. The instructions are intended to require an examinee to indicate which picture represents what he would have to see in order to verify *his* inference; Molitor (1972) states that "A correct choice on the verification item is independent of a correct choice on either or both parts of the inference item. A score of +1 is given for any choice that represents the observation that would be required to verify the examinee's inference" (administration manual, p. 164). However, the directions do not state that the inference to be verified is the one made by the examinee. Instead, the directions allow several responses to qualify as correct. For example, in the sample test item reproduced here, the directions for the verification item state: "Which one of all these six pictures shows what you would have to see in order to be sure how the window was broken?" *Any* of the top three pictures could establish how the window was broken, regardless of what was originally inferred as the cause. If responses to verification items are to be evaluated in terms of the responses to inference items, then the logical connection between the two should probably be made explicit.

Finally, Molitor notes that a number of test items presented situations familiar to examinees, and concluded that examinees sometimes responded on the basis of their familiarity with the events or relationships pictured, rather than deriving their responses from a consideration of the data provided. This insight raises questions about the validity of the instrument as a measure of any sort of reasoning process, and might imply that some test items fail to sample the intended abilities.

Despite these difficulties, however, the Science Process Skills Test appears promising. It must be remembered that the instrument is still in its developmental stages; the weaknesses identified in this review should be considered more as suggested revisions than as deficiencies of a finished product. In general, the Science Process Skills Test would appear to merit further refinement and development.

Additional information:

Molitor, L. L. Development of the science process skills test. Paper presented at the annual meeting of the National Association for Research in Science Teaching, April, 1972.

Molitor, L. L., & George, K. D. Development of a test of science process skills. *Journal of Research in Science Teaching*, 1976, *13*, 405-412.

Test citation:

Molitor, L. L. Science Process Skills Test. Unpublished test. Towson, Maryland: Physics Department, Science Education Group, Towson State College, 1972.

SMITH-STURGEON CONDITIONAL REASONING TEST

Author(s): Edward L. Smith and Joanne Sturgeon
Year of publication: unpublished
Most recent edition: 1969

Description:

The Smith-Sturgeon Conditional Reasoning Test was developed as an evaluation instrument as part of the Cornell Critical Thinking Readiness Project (phase II C). The test is intended to measure mastery of four conditional reasoning principles:[1]

(1) particular inversion (denying the antecedent; an invalid move)
(2) particular conversion (affirming the consequent; an invalid move)
(3) particular contraposition (denying the consequent; valid)
(4) particular transitivity (valid)

The format of the test is considerably different from the usual paper-and-pencil, multiple-choice type of instrument. The test is administered individually. In the first part of the test, the conditional reasoning principles are illustrated through the use of simple chemical reactions demonstrated to the examinee as the test is administered.[2] In the second part of the test, a specially designed model house is used to illustrate conditional relations between switches, handles, a bell and a light. The test consists of 24 items, equally divided among the four principles. For a group of six items testing for mastery of a single principle, three items are presented in suppositional form (the examinee is asked to suppose that the minor premise in the conditional reasoning argument is true). The remaining three items are given in factual form (the truth of the minor premise is demonstrated to the examinee). Examinees must demonstrate mastery of the major premise in each part of the test. For the chemicals part, the conditional statement, "If vinegar is added to [baking] soda, then the soda bubbles" serves as the major premise. For the house part, the conditional statement, "If the big handle is up, then the bell will ring" serves as the major premise. The major premise in each section is demonstrated to the examinee until the examinee can successfully predict the result of adding vinegar to soda or of raising the big handle. In addition, the major premise for each section is reinforced visually by

---

[1] The logical principles discussed in this review are classified by the system proposed by Ennis (1976).

[2] The first part of the Smith-Sturgeon Conditional Reasoning Test served as the model for the development of the Illinois Conditional Reasoning Test, also reviewed here. The reader is referred to the review of the Illinois Conditional Reasoning Test for purposes of comparison; the description of that instrument should acquaint the reader with the major features of the first part of the Smith-Sturgeon Conditional Reasoning Test.

Description (continued):

the use of a picture board and interchangeable drawings which represent the relevant conditional relationships. Examinees who fail to master the major premise for a part of the test cannot complete that part.

Scoring is somewhat complicated, and requires the use of a special scoring form. A number of part scores as well as a total test score can be computed. Examinees must provide both a correct response and an adequate justification in order to receive credit on any item. Scoring is by the 'rights only' method. From the examinee's perspective, test directions are fairly simple and straightforward, requiring only that the examinee respond to oral questions. The examiner, however, must read the script supplied with the instrument, manipulate the appropriate test materials, and possibly record and score the responses as well.

The Smith-Sturgeon Conditional Reasoning Test was used with students in grades 1-3, but might also be appropriate for older examinees, providing that they do not have a chemistry background which might influence scores on the chemicals part of the test. Although no manual accompanies the instrument, Ennis et al. (1969) serves this purpose, and includes information on the construction, administration and scoring of the instrument, as well as reliability and validity information.

Sample test item (house part):

[Capitalized words are those spoken by the examiner. Quoted words are responses or possible responses given by the examinee.]

11. IF THE BIG HANDLE IS UP, WHAT DO YOU KNOW ABOUT THE BELL?

[checks mastery of the major premise: If the big handle is up, the bell will ring.]

12. PRETEND THAT YOU PUSHED THE BUTTON ON THE BELL AND THE BELL RANG. WHAT ARE YOU PRETENDING?

"I'm pretending that I pushed the button on the bell and the bell rang."

[checks understanding of the problem situation]

13. IF YOU PUSHED THE BUTTON ON THE BELL AND THE BELL RANG, WOULD YOU SAY THAT THE BIG HANDLE IS UP OR ISN'T UP OR WOULD YOU SAY THAT MAYBE IT IS AND MAYBE IT ISN'T?

[checks mastery of particular conversion principle for suppositional item]

14. HOW WOULD YOU KNOW THAT?

[requests justification for response given in 13]

Reliability and validity information:

Ennis et al. (1969) present reliability and validity data for this instrument. Kuder-Richardson 21 reliability estimates for part and total test scores are presented for each grade level included in the study, and ranged from r = .07 (contraposition principle score; grade 3) to r = .95 (fallacy items score; grade 2). Total test score reliability for all grades combined was estimated as r = .84. Means and standard deviations for part and total test scores at each grade level are also reported.

Ennis et al. (1969) also argue that the content validity of the measure is supported by an examination of the test scripts which describe and illustrate the principles of conditional reasoning whose mastery is to be assessed. Further, the correct response to each test item is implied by the principle tested for in that item. The construct validity of the Smith-Sturgeon Conditional Reasoning Test was weakened somewhat by a failure to find higher correlations for identical traits measured by different methods than for different traits measured by either the same or different methods. The multitrait-multimethod matrix presented considers the four conditional reasoning principles as traits and the two test parts as different methods of measurement. However, construct validity of the instrument was supported by the finding that correlations between part scores of the two validity principles and the two invalidity principles were higher than the correlations between part scores for a valid principle and an invalid principle. This result was consistent with the theoretical framework outlined by the authors. In addition, correlations of total test score with I.Q. (as measured by the Wechsler Intelligence Scale for Children (WISC)); verbal ability (as measured by the verbal subscore on the WISC), socio-economic status, chronological age and sex generally conformed to theoretical expectations. The authors report a correlation of r = .51 between total test score and verbal subscore on the WISC, and a correlation of r = .57 between total test score on the Smith-Sturgeon Conditional Reasoning Test and WISC total score. Both correlations are based on the data from combined grades. For all grades combined, total test score correlated r = .44 with socio-economic status, r = .19 with chronological age, and r = -.13 with sex. Correlations were based on test administration to groups of children having no instruction in logic.

Item discrimination and difficulty indices are also presented.

Availability:

The test is unpublished, but is presented and discussed extensively in Ennis et al. (1969). This work also describes the research project for which the test was developed, and is available on ERIC (# ED 040 437).

Comments:

The Smith-Sturgeon Conditional Reasoning Test represents an interesting and imaginative approach to assessing mastery of four conditional reasoning principles. The format seems likely to prove interesting to examinees, and the efforts to insure that examinees understand basic problem information should result in a more accurate reflection of individual abilities. The requirement that examinees justify their responses provides additional evidence from which mastery or non-mastery of a particular conditional reasoning principle may be inferred.

For the purposes of the study in which this instrument was used, mastery of a logical principle was operationally defined as at least five correct responses (including adequate justification) on the six items measuring a particular principle. Three or fewer correct responses for a single principle is given as the criterion for probable non-mastery of that principle. No mastery statement is made for a score of four correct. The numerous part scores which can be computed (10) should be helpful in identifying specific weaknesses in conditional reasoning ability.

It would be helpful, however, to have test-retest reliability estimates for both part and total test scores. The reported Kuder-Richardson 21 reliability estimates provide an indication of the internal consistency of the subtests and the test as a whole, but do not reflect the consistency of measurements over time. Some indication of the instrument's dependability as a measure of conditional reasoning ability would be of value to prospective test-users.

Probably the most serious drawback to the Smith-Sturgeon Conditional Reasoning Test is the complicated administration and scoring procedures. The test developers' efforts to make scores on the instrument meaningful indicators of conditional reasoning ability have resulted in a measure possessing, in the judgment of this reviewer, a high degree of validity. But the individual administration, the extensive materials required in order to administer the test, and the scoring procedures which rely on the informed judgment of the tester in evaluating responses are likely to discourage potential test-users. The extra effort required in administering and scoring the test, however, should be evaluated against the potential accuracy of appraisal provided by the instrument. And although the Smith-Sturgeon Conditional Reasoning Test may not receive extensive use, it represents a promising line for future test development; test developers may find that minor modifications will incorporate the strengths of the measure in their own instruments while minimizing its drawbacks.

Additional information:

Ennis, R. H., Finkelstein, M. R., Smith, E. L., & Wilson, N. H. Conditional logic and children. Albany, New York: New York State Education Department, Division of Research, 1969.

Test citation:

Smith, E. L., & Sturgeon, J. Smith-Sturgeon Conditional Reasoning Test. Unpublished test. Albany, New York: New York State Education Department, Division of Research, 1969.

Author(s): Committee on Evaluation Materials, Institute for Propaganda Analysis
Year of publication: 1938 (copyright)
Most recent edition: 1938

Description:

The Strong and Weak Argument Test consists of a series of 10 questions, each followed by a number of short arguments (usually one sentence). For each argument presented, examinees must decide if it is strong and important, or weak and unimportant. Arguments on both sides of each question are given. Examinees are asked to assume that each argument presented is true, and are required to evaluate 40 arguments in the course of the test.

No manual is provided with the test, and hence, details of test administration, scoring, score interpretation and reliability and validity information are unavailable. No caution against guessing is specified in the test directions, which appear on the first page of the instrument. A space for 'total score' is provided on the front of the test booklet; it seems likely that only a single score for this measure is computed, and that such a score consists of the total number of correct responses. A statement on page 1 of the test suggests that the test is intended for use at the high school level and above.

Sample test item:

[The following directions to examinees are excerpted from the first page of the test]:

> If you think it is a strong, important argument, well worth considering, draw a circle around the word Strong. Whenever you feel the argument is weak, and unimportant, draw a circle around the word Weak. It makes no difference whether the argument is on the side of the question with which you agree or not. Consider only whether it is a strong or weak argument, regardless of whether or not you personally agree with it. Be sure you mark every argument as being comparatively Strong or comparatively Weak.

[emphasis, capitalization in original]

IV. Are labor unions desirable in the United States?

Strong Weak 1. Labor unions force men to go out on strike and these strikes often cause a lot of trouble.

Strong Weak 2. Labor unions are the only form of organization which will adequately protect the interests and rights of the workers.

etc.

Reliability and validity information:

No reliability or validity information is currently available for this test.

Availability:

The test was copyrighted by the Institute for Propaganda Analysis, New York, New York, but apparently never published.

Comments:

The Strong and Weak Argument Test, in the judgment of this reviewer, appears to have some face validity as a measure of the ability to evaluate the strength of arguments. The directions to examinees seem clear and straightforward, and the language used appears appropriate for high school students and above. It is unfortunate that reliability and validity information for this test is unavailable, since such information would aid greatly in evaluating this instrument.

At least one difficulty with the measure can be identified by inspection, however. Because only two alternative responses are provided for each item, the probability of chance success is fairly high (50%). On the assumption that the total test score is computed as the total number of correct responses, a high score on the test might reflect lucky guessing as much or more than it would indicate proficiency in evaluating the strength of arguments.

Further, the authors of the test might easily have constructed the instrument to reveal additional information about an examinee's performance. For example, examinees might be required to determine two things for each argument: (1) the direction of support provided by the argument for the question being addressed; and (2) the degree of support provided by the argument. The question posed for evaluation in each test item might be rephrased as a declarative statement, in order that examinees could judge whether a given argument would provide evidence against or evidence in favor of the statement. A further refinement might be produced by incorporating irrelevant arguments as alternatives. For arguments marked as irrelevant, examinees would not be required to evaluate the direction of support provided. It seems likely that these comparatively simple revisions might result in a more accurate appraisal of an examinee's ability to evaluate arguments by sampling additional abilities important to appraising arguments critically.

One major difficulty with this instrument is the fact that it appears to test for the values of the test-taker. Someone with conservative values, for example, might well judge IV-1 to be a strong argument, while someone with liberal values might judge it to be a weak argument.

Comments (continued):

Another major difficulty with this instrument is the lack of information available concerning it. In the opinion of this reviewer, the Strong and Weak Argument Test shows considerable promise, and is probably useful as a source of ideas for constructing instruments to assess at least some aspects of the ability to evaluate arguments.

Test citation:

Committee on Evaluation Materials. Strong and Weak Argument Test. Unpublished test. New York: Institute for Propaganda Analysis, Incorporated, 1938.

Author(s): unknown
Year of publication: unpublished
Most recent edition: unknown

Description:

The Test of Ability to Find Appropriate Illustrations consists of 20 items. Each item is a statement which might support, question or be irrelevant to one or more of five generalizations listed at the beginning of the test. For each statement, examinees must indicate the relation of the statement to each of the five generalizations given at the beginning of the test (in effect, five responses are required for each test item).

No manual is provided with the test, and hence details of test administration, scoring, score interpretation, and reliability and validity information are unavailable. The instrument itself provides no clues about the number and types of scores which might be computed or the method of their computation. Test directions contain no cautions against guessing. A sample problem is given on the first page of the test. In the judgment of this reviewer, the language used in the instrument appears appropriate for students at or above grade 10.

Sample test item:

[The following directions to examinees are excerpted from the first page of the test]:

> In the blank beside each statement, indicate with an X the relation of that statement to each of the generalizations above. The column headed "S" means "supports the generalization." "Q" means "questions or casts doubt upon the generalization." "I" means "irrelevant to the generalization."
>
> Note that the criterion for marking an item "S" for "Q", rather than "I" is not merely -- "has this item any bearing whatsoever upon one of the above generalizations?" Rather, ask the question, -- "could this item reasonably be used in a discussion as evidence for or against one of these generalizations?

[emphasis in original]

[The following generalizations are given at the beginning of the instrument]:

1. CUSTOMS AND BELIEFS TEND TO PERSIST LONG AFTER THE CONDITIONS THAT GAVE RISE TO THEM HAVE CEASED TO OPERATE.

2. BUSINESS MEN BENT UPON PROFIT HAVE BEEN THE MEANS OF EXTENDING A BETTER WAY OF LIFE OVER WIDE AREAS.

Sample test item (continued):

3. SITUATIONS THAT THREATEN WAR GROW OUT OF THE ACTIVITIES OF BUSINESS AND COMMERCE.

4. EVEN WITHIN THE THINKING OF A SINGLE INDIVIDUAL, TWO OR MORE CHERISHED VALUES ARE OFTEN THROWN INTO CONFLICT.

5. IN THE UNITED STATES, IMPORTANT ISSUES ARE DECIDED IN ACCORDANCE WITH THE DESIRES OF A MAJORITY OF THE ADULT POPULATION.

[capitalization in original]

	S	Q	I
1.			
2.			
3.			
4.			
5.			

17. Although no reputable psychologist any longer supports the doctrine of faculty psychology, courses purporting to "train the will or the memory" continue to flood what appears to be a wide market.

Reliability and validity information:

No reliability or validity information for this test is currently available.

Availability:

The test was apparently never published.

Comments:

Evaluation of the Test of Ability to Find Appropriate Illustrations is hampered considerably by the lack of information concerning this measure. However, several comments can be made based upon an inspection of the instrument.

The title of this test seems somewhat misleading. Inspection of both the test items and directions to examinees suggests to this reviewer that the instrument might be more accurately described as an attempt to assess two aspects of critical thinking ability: (1) judging whether a statement is relevant to a given generalization; and (2) judging the direction of support provided by the statement for the generalization. Also, this measure potentially might provide considerable information about these abilities of examinees. For example, the statement given in the sample test item is clearly relevant to generalization 2 (which it appears to question or cast doubt upon). An examinee who consistently identified statements as irrelevant to particular

Comments (continued):

generalizations when in fact such statements constituted counter-evidence to the generalization may have difficulty in judging relevance. Alternatively, a comparison of an examinee's frequency of labelling counter-evidence statements as irrelevant with the frequency of labelling supporting evidence as irrelevant might reveal that this particular examinee has difficulty in recognizing the relevance of statements only when such statements serve as counter-evidence to the generalization in question. In addition, a number of other 'scores' might be computed, which reflected specific difficulties in judging the relevance of statements (number of supporting statements judged as casting doubt upon particular generalizations might provide an indication that an examinee had difficulty in distinguishing the direction of support provided by a statement for a generalization, for example). However, in the absence of any information concerning the construction, scoring and interpretation of this test, the instrument should probably be viewed as promising but incomplete.

An unfortunate feature of the test booklet is that the directions to examinees and the five generalizations of concern appear only on page 1. Unless examinees memorize the generalizations, they will be forced to return to the first page of the booklet in order to recall the generalizations with which they are concerned. This arrangement of the test is inconvenient, and may also introduce distractions which would lower the validity of the instrument. A simple solution to this problem might be to reproduce the generalizations (and possibly the explanation of the alternative responses) at the top of each page.

Finally, the intended answers for each item were not always clear to this reviewer. For example, in the sample test item reproduced here, the statement may or may not be relevant to generalization 3, depending upon the interpretation made. The statement might be taken as an example of an "activity of business and commerce" which did not threaten war, and hence, should be marked as casting doubt upon generalization 3. Further, the statement may also support generalization 1 on the assumption that those offering courses to "train the will or the memory" accept the truth of faculty psychology. Without this assumption, however, the statement appears irrelevant to generalization 1. Similar problems arise in other test items, which suggests that the test could profit from a logical analysis of both the items and the intended answers.

In general, the Test of Ability to Find Appropriate Illustrations is far from a finished product, although it does show some promise. The ability to assess the relevance of statements for a generalization is an important aspect of critical thinking ability which often is ignored in general tests of critical thinking ability. This instrument provides a beginning, but only a beginning, for assessing this aspect of critical reasoning.

Test citation:

(Author unknown). Test of Ability to Find Appropriate Illustrations. Unpublished test. No further information available.

Author(s): William V. Haney
Year of publication: 1955
Most recent edition: 1975

Description:

The Uncritical Inference Test, according to a statement on the test booklet, is "...designed to determine [examinees'] ability to think accurately and carefully" (emphasis in original). Examinees are presented with three short stories in the course of the test. A series of statements follows each story. For each statement, the directions require examinees to determine whether it is:

a. "T" -- meaning: On the basis of the information presented in the story the statement is DEFINITELY TRUE.

b. "F" -- meaning: On the basis of the information presented in the story the statement is DEFINITELY FALSE.

c. "?" -- meaning: The statement MAY be true (or false) but on the basis of the information presented in the story you cannot be definitely certain. (If any part of the statement is doubtful, mark the statement "?").

[emphasis, capitalization in original]

Examinees evaluate a total of 75 statements. A sample problem is provided at the beginning of the instrument. Test directions state that the information presented in each story is to be taken as "definitely accurate and true" (emphasis in original). The directions also prohibit examinees from returning to previous questions to revise or complete answers. No time limit is specified.

A three-page report of data relevant to the Uncritical Inference Test is supplied with the instrument. The report contains primarily reliability and validity information, and provides no guidelines for test administration, scoring, or score interpretation. A short list of related references (all authored by Haney) and an answer key are included, however.

Sample test item:

STORY C

A businessman had just turned off the lights in the store when a man appeared and demanded money. The owner opened a cash register. The contents of the cash register were scooped up and the man sped away. A member of the police force was notified promptly.

Statements About Story C

1. A man appeared after the owner had turned off his store lights. T F ?

2. The robber was a man. T F ?

3. The man who appeared did not demand money. T F ?

4. The man who opened the cash register was the owner. T F ?

....

18. The following events occurred: someone demanded money, a cash register was opened, its contents were scooped up, and a man dashed out of the store. T F ?

Reliability and validity information:

Haney presents considerable reliability and validity data for earlier versions of the Uncritical Inference Test. He states, however, that the present form is superior to the earlier versions because "...it is longer (76 [sic] items) and the items are better discriminators" (data sheets, p. 1). Two comparable 67 item tests were developed prior to the present version of this instrument. Split-half reliability for Form A is reported as r = .762, and Form B split-half reliability is given as r = .818. Stepped-up reliabilities (by Spearman-Brown) for the full-length tests yielded values of r = .928 (Form A) and r = .947 (Form B). Test-retest reliability coefficients for both forms are also reported. For a sample of N = 164 Northwestern undergraduates, a test-retest reliability coefficient of r = .865 (Form A) was obtained. A sample of N = 136 Northwestern undergraduates yielded a test-retest coefficient of r = .900 (Form B). Additional correlations are reported for six samples of Northwestern undergraduates. Three groups received instruction between administration of Form A and Form B; the remaining three groups did not. Reported correlations ranged from r = .559 (Form B -- no training -- Form A; N = 42) to r = .739 (Form A -- training -- Form B; N = 61).

Haney also notes that test items were judged valid by a "...board of qualified observers" (data sheets, p. 2). For the three groups receiving instruction, posttest scores were found to be significantly higher than

Reliability and validity information (continued):

pretest scores ($p < .001$). This difference was still significant when practice effect, computed from improvement in posttest scores of the uninstructed groups was subtracted from posttest scores of the instructed groups ($p < .01$). Additional correlations are reported for Form A or Form B of the Uncritical Inference Test and scores on other measures. Of particular interest is the correlation between scores on Form A and total test score on the Watson-Glaser Critical Thinking Appraisal. For a sample of $N = 115$, the correlation obtained was $r = .249$ (Watson-Glaser administered first). Correlations of Form A with the Inference subtest of Watson-Glaser ranged from $r = .092$ ($N = 115$; Watson-Glaser administered first) to $r = .215$ ($N = 47$; Form A administered first). The form of the Watson-Glaser used was not specified.

Availability:

The test is currently in print and is available from William V. Haney Associates, 2453 Cardinal Lane, Wilmette, Illinois, 60091.

Comments:

The Uncritical Inference Test suffers from a number of serious difficulties. The most important of these is the test items themselves. Approximately two-thirds of the answers are keyed as "?" (doubtful), which suggests that a pathological doubter would do well on this test without thinking critically at all. And examination of the test items further suggests that the justification for the keyed answers is often suspect. For example, in the sample test item reproduced here, the keyed answers are as follows: ?, ?, F, ?, ?. Statement 1 is doubtful presumably because the story does not state specifically that the owner turned off the lights (a "businessman" did so). Statement 2 is doubtful because the story does not state that a robbery occurred, only that money was demanded. Statement 3 is clearly false since it contradicts information given in the story. Statement 4 is doubtful since the gender of the owner is not specified. And finally, statement 18 is doubtful, presumably because the story does not state specifically that a man dashed out of the store (he "sped away", however). In the opinion of this reviewer, these sorts of items are misleading. Rather than reflecting the degree to which examinees are able to make critical inferences, or even uncritical inferences, the test seems more to reflect the willingness of examinees to make any inferences whatsoever beyond the information explicitly presented in the story. This instrument might better be referred to as a critical un-inference test.

Other difficulties also undermine the validity of this measure. The test directions which prohibit returning to items (for any reason) once they have been passed effectively penalize examinees who discover the nature of the test. The low reported correlations between scores on the Uncritical

Comments (continued):

Inference Test and scores on other measures lead Haney to conclude that the test is unique. Unfortunately, however, the measure appears to differ even from other measures with which it could be expected to correlate highly (such as the Inference subtest of the Watson-Glaser Critical Thinking Appraisal). Further, no validity or reliability information whatsoever is presented for the current form of the test; the reported correlations were obtained for earlier versions of the instrument. Hence, the validity and reliability of this measure appear far from established, and it is unclear what sort of evidence the reported correlations are intended to provide for the reliability and validity of the current instrument.

In general, the Uncritical Inference Test seems far from a finished product. Although the attempt to measure the tendency to make inferences uncritically is commendable, the reliability and validity of this instrument have yet to be established. Perhaps then scores on this measure will admit of meaningful interpretation.

Test citation:

Haney, W. V. Uncritical Inference Test. Wilmette, Illinois: William V. Haney Associates, 1955.

Author(s): Goodwin Watson and Edward M. Glaser
Year of publication: 1951
Most recent edition: 1964

Description:

The Watson-Glaser Critical Thinking Appraisal is probably the most widely used critical thinking test currently available. It is designed to measure examinees' skills in five aspects of critical thinking: (1) inference (20 items) -- ability to discriminate among degrees of truth or falsity in inferences drawn from data; (2) assumption recognition (16 items) — ability to recognize unstated assumptions and presuppositions which are taken for granted in statements; (3) deduction (25 items) — ability to reason in accord with the principles of class and conditional logic; (4) interpretation (24 items) -- ability to judge whether given generalizations are reasonably supported by the data; and (5) evaluation of arguments (15 items) -- ability to discriminate between strong, relevant arguments and weak or irrelevant arguments. The instrument is available in two forms (Ym and Zm), which replace the earlier forms Am and Bm. Each form consists of 100 items.

The instrument is composed of five subtests, each subtest dealing with one of the specific aspects of critical thinking ability identified above. Each subtest is preceded by a sample item. The inference subtest presents examinees with a series of short passages. The information contained in the passages is to be taken as factual. Several inferences follow each passage. For each inference, examinees must decide if it is true, probably true, probably false, false, or that no decision is possible due to insufficient data. The assumption recognition subtest consists of several statements, each of which is followed by several proposed assumptions. For each proposed assumption, examinees must decide if the speaker, in making the statement provided, is taking the proposed assumption for granted (justifiably or not). In the deduction subtest, examinees are given two premises and must decide which (if any) of several conclusions follows necessarily from the premises. The interpretation subtest provides examinees with a series of passages for study. Each passage is followed by a list of statements. For each statement, examinees must decide if it follows beyond a reasonable doubt from the information presented in the passage, or if it does not follow. The final subtest, evaluation of arguments, is composed of several questions, each followed by a list of short arguments (usually on both sides of the question). Examinees must rate each argument as strong or weak. Arguments are meant to be taken as "true".

Part scores and total test scores may be computed; however, the authors advise against using part scores as measures of individual achievement in specific critical thinking areas (manual, p. 9). Scoring is by the 'rights only' method. A time limit of 50 minutes is suggested for completion, but the authors note that examinees should be given every opportunity to complete the test, even if the time required exceeds 50 minutes. The subtests are each timed separately (but not equally); examinees who fail to complete a subtest within the time limit may return to finish at the conclusion of the test.

Description (continued):

A manual is provided with the test, and contains reliability and validity information, discussions of the development of the test, rationale, scoring procedures and test administration, norms, and a brief treatment of the concept of critical thinking on which the test is based. No specific statement is made concerning the age groups for which the test is appropriate.

Sample test item (argument evaluation):

Should groups in this country who are opposed to some of our government's policies be permitted unrestricted freedom of press and speech?

98. Yes; a democratic state thrives on free and unrestricted discussion, including criticism.

99. No; the countries opposed to our form of government do not permit the free expression of our point of view in their territory.

100. No, if given full freedom of press and speech, opposition groups would cause serious internal strife and make our government basically unstable, eventually leading to loss of our democracy.

Reliability and validity information:

The test manual presents considerable reliability and validity data, as well as norms for grades 9-12, grades 9-12 combined, college senior women, and liberal arts college freshmen. Means and standard deviations (and in some cases standard errors) are also provided for a number of other special groups, in addition to being provided for the normative samples. The manual also includes tables which relate scores on Form Am to those on the current Ym and Zm forms.

Odd-even reliability estimates, corrected by Spearman-Brown, are presented for each group on which the test was normed. Coefficients range from r = .77 (Form Zm, grade 9) to r = .87 (Form Ym, grade 12). Individual subtest reliabilities for each form are also given. Odd-even reliability estimates, corrected by Spearman-Brown, vary between r = .41 (deduction subtest, Form Zm) and r = .74 (assumption recognition subtest, Form Ym). These reliability estimates were derived from the grade 10 normative sample (N = 2947, Form Ym; N = 2995, Form Zm). Item difficulty information and item discrimination values are also presented for this group.

The authors suggest that the subject matter of critical thinking is not clearly defined, and therefore that the content validity of the Watson-Glaser Critical Thinking Appraisal should be evaluated in the instructional context in which the measure will be used. Construct validity of the measure is claimed on the basis of low subtest inter-correlations (for grade 10

Reliability and validity information (continued):

normative sample). Watson and Glaser conclude that the generally low inter-correlations (ranging from r = .21 to r = .50) "...support the contention that relatively distinctive abilities are being measured with sufficient overlap to warrant their inclusion in one total score" (manual, p. 14). Two factor-analytic studies are also cited as evidence for the construct validity of the measure. Finally, construct validity is also supported by subtest-total test score correlations which range from r = .56 (argument evaluation, Form Zm) to r = .79 (interpretation subtest, Form Ym). These correlations are also based upon the grade 10 normative sample. The authors note, however, that subtest-total test correlations are slightly inflated due to the correlation of the subtest with itself being included in the subtest-total test score correlation.

No evidence is offered for the predictive validity of the instrument. Instead, the manual suggests that the test's predictive validity should be established empirically in each situation in which the measure will be used.

Availability:

The test is currently in print and available from Harcourt, Brace, Jovanovich, Inc., 757 Third Avenue, New York, New York, 10017.

Comments:

The Watson-Glaser Critical Thinking Appraisal appears to be a potentially useful instrument for assessing the critical thinking ability of examinees. Although the instrument does not sample every aspect of critical thinking ability one might wish (judging reliability of observation statements, judging authorities, and semantical skills are notably absent), the test does include important aspects of critical thinking ability in those sampled. The manual is complete and generally informative; test directions are clear and administration is relatively uncomplicated. The ease of administration, coupled with the amount of information potentially provided by the test, should make this measure particularly attractive to prospective users.

However, the test is not without its problems. The evaluation of arguments subtest includes items which are clearly value-laden; examinees who do not share the value orientation of the test authors may be penalized. For example, in the sample test item reproduced here, statements 98 and 100 are keyed as strong arguments; statement 99 is keyed as a weak argument. Someone who believes that a limited democratic form of government is not optimal for this country might decide that statement 100 was a weak argument. Many of the items in this section appear to evaluate the value systems of

Comments (continued):

examinees in addition to their ability to evaluate arguments. Further, this reviewer feels that the subtest could be improved by eliminating the general pro-America bias which seems to pervade the section. Although the inclusion of controversial topics may be beneficial, the keyed answers may reflect the values of the test constructor as much as an accurate appraisal of the strength of the argument.

The inference subtest items keyed as "probably true" or "probably false" also present difficulties. More sophisticated examinees might (justifiably) judge that insufficient evidence has been provided on which to base a decision. They would therefore be penalized for recognizing the limitations of the information provided in the passage. The directions for this subtest also make it clear that examinees are to rely on "common knowledge" in evaluating the statements. It seems likely that the nature and extent of "common knowledge" could vary considerably among examinees, which might distort scores obtained on this subtest.

The assumption recognition subtest directions encourage examinees to look for assumptions which are necessarily taken for granted. Unfortunately, the nature of this necessity is left unspecified. Presuppositions might be construed as logically necessary assumptions, but assumptions which serve as missing premises in arguments are not necessary in this sense (see Ennis, 1961). However, the majority of proposed assumptions are not assumptions in either of the two senses identified above. Rather, they are beliefs which might be reasonably attributed to the speaker of the statement (similar to Gricean conversational implicatures). The confusion of at least these three senses of "assumption" tends to obscure the exact nature of the ability purportedly measured by this subtest.

In general, the Watson-Glaser Critical Thinking Appraisal is a test with a long history of use. The measure is one of the few currently available which provide a general assessment of critical thinking ability rather than concentrating on a single aspect. The manual contains sufficient information to enable prospective users to make an informed evaluation of the instrument (although more current norms and greater attention to test validity would be a welcome addition). In the judgment of this reviewer, the test is one of the better available, despite the limitations and difficulties noted earlier.

Additional information:

Broadhurst, N. A. An item analysis of the Watson-Glaser critical thinking appraisal (form Ym). *Science Education*, 1970, *54*(2), 127-132.

Buros, O. K. *Third mental measurements yearbook (1949)*. Reviews appear on p. 570.

Buros, O. K. *Fifth mental measurements yearbook (1959)*. Reviews appear on pp. 796-799.

Buros, O. K. *Seventh mental measurements yearbook (1972)*. Reviews appear on pp. 1214-1215.

Burton, A., & Walther, J. Adult norms for the Watson-Glaser tests of critical thinking. *Journal of Psychology*, 1945, *19*, 43-48.

Chang, E. C. Norms and correlates of the Watson-Glaser critical thinking appraisal and selected variables for Negro college students. Doctoral dissertation, University of Oklahoma, 1969.

Ennis, R. H. An appraisal of the Watson-Glaser critical thinking appraisal. *Journal of Educational Research*, 1958, *52*(4), 155-158.

Ennis, R. H., & Macmillan, C. J. B. *Critical thinking reviews* (second edition). Ithaca, New York: Cornell Critical Thinking Project, 1968.

Follman, J., Miller, W., & Burg, E. Statistical analysis of three critical thinking tests. *Educational and Psychological Measurement*, 1971, *31*(2), 519-520.

Gibson, J. W., Kibler, R. J., & Barker, L. L. Some relationships between selected creativity and critical thinking measures. *Psychological Reports*, 1968, *23*, 707-714.

Glaser, E. M. An experiment in the development of critical thinking. Doctoral dissertation, Columbia University, 1941.

Helm, C. R. Watson-Glaser-DAT graduate norms. Master's thesis, University of Toledo, 1967.

Kooker, E. W. The relationship between performance in a graduate course in statistics and the Miller analogies test and the Watson-Glaser critical thinking appraisal. *Journal of Psychology*, 1971, *77*(2), 165-169.

Rust, V. I. Factor analyses of three tests of critical thinking. Doctoral dissertation, University of Illinois, 1959.

Rust, V. I. Factor analyses of three tests of critical thinking. *Journal of Experimental Education*, 1960, *29*, 177-182.

Additional information (continued):

Shatin, L., & Opdyke, D. A critical thinking appraisal and its correlates. *Journal of Medical Education*, 1967, *42*, 759-792.

Troxel, V. A., & Snider, C. Correlations among student outcomes on the test of understanding science, Watson-Glaser critical thinking appraisal, and the american chemical society cooperative examination--general chemistry. *School Science and Mathematics*, 1970, *70*(1), 73-76.

Test citation:

Watson, G., & Glaser, E. M. Watson-Glaser Critical Thinking Appraisal. New York: Harcourt, Brace, Jovanovich, Inc., 1951, 1964.